PCT International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12P 19/34, C07H 21/04 — A1

(11) International Publication Number: WO 99/23256

(43) International Publication Date: 14 May 1999 (14.05.99)

(21) International Application Number: PCT/US98/23168

(22) International Filing Date: 30 October 1998 (30.10.98)

(30) Priority Data:
60/064,358 30 October 1997 (30.10.97) US

(71) Applicant *(for all designated States except US)*: COLD SPRING HARBOR LABORATORY [US/US]; Bungtown Road, Cold Spring Harbor, NY 11724 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: WIGLER, Michael [-/US]; 52 Snake Hill Road, Cold Spring Harbor, NY 11724 (US). LUCITO, Robert [-/US]; 41 Liberty Avenue, Mineola, NY 11501 (US).

(74) Agents: BALDWIN, Geraldine, F. et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: PROBE ARRAYS AND METHODS OF USING PROBE ARRAYS FOR DISTINGUISHING DNA

(57) Abstract

It is an object of the present invention to provide a solution to problems associated with the use of microarray technology for the analysis DNA. The present invention provides compositions and methods for the use of simple and compound representations of DNA in microarray technology. The present invention is also directed to methods for the production of High Complexity Representations (HCRs) of the DNA from cells.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

PROBE ARRAYS AND METHODS OF USING PROBE ARRAYS FOR DISTINGUISHING DNA

This invention was made with Government support under Contract Nos. 5R35 CA 39829-13 and 5P50 CA 68425-03 awarded by the National Institutes of Health. The Government has certain rights to this invention.

This application claims the benefit of U.S. Provisional Application No. 60/064,358, filed 30 October, 1997.

1. FIELD OF THE INVENTION

The field of the invention is genetic analysis.

2. BACKGROUND OF THE INVENTION

2.1. MICROARRAY TECHNOLOGY

Although global methods for genomic analysis, such as karyotyping, determination of ploidy, and more recently comparative genomic hybridizaton (CGH) (Feder, et al.,1998, Cancer Genet.Cytogenet., 102:25-31; Gebhart, et al., 1998, Int.J.Oncol. 12:1151-1155; Larramendy, et al., 1997, Am.J.Pathol., 151:1153-1161; Lu, et al., 1997, Genes Chromosomes Cancer, 20:275-281, all of which are incorporated herein by reference) have provided useful insights into the pathophysiology of cancer and other diseases or conditions with a genetic component, and in some instances have aided diagnosis, prognosis and selection of treatment, current methods do not afford a level of resolution of greater than can be achieved by standard microscopy, or about 5 - 10 megabases. Moreover, while many particular genes that are prone to mutation can be used as probes to interrogate the genome in very specific ways (Ford, et al., 1998, Am.J.Hum.Genet., 62:676-689; Gebhart, et al., 1998, Int.J.Oncol., 12:1151-1155; Hacia, et al., 1996, Nat.Genet., 14:441-447, all of which are incorporated herein by reference), this one-by-one query is an inefficient and incomplete method for genetically typing cells.

With the advent of microarray, or "chip" technology, it is now clearly possible to contemplate obtaining a high resolution global image of genetic changes in cells. Two general approaches can be conceived. One is to profile the expression pattern of the cell using microarrays of cDNA probes (DeRisi, et al., 1996, Nat.Genet., 14:457-460). This method is very likely to yield useful information about cancer, but suffers limitations. First, the interpretation of the data obtained and its correlation with disease process is likely to be a complex and difficult problem: multiple changes in gene expression will be observed that are not relevant to the disease of interest. Second, our present cDNA collections are not complete, and any chip is likely to be obsolete in the near future. Third, while a picture of the current state of the cell might be obtained, there would be little direct information about how the cell arrived at that state. Lastly, obtaining reliable mRNA from biopsies is likely to be a difficult problem, because RNA is very unstable and undergoes rapid degradation due to the presence of ubiquitous RNAses.

The second approach is to examine changes in the cancer genome itself. DNA is more stable than RNA, and can be obtained from poorly handled tissues, and even from fixed and archived biopsies. The genetic changes that occur in the cancer cell, if their cytogenetic location can be sufficiently resolved, can be correlated with known genes as the data bases of positionally mapped cDNAs mature. Thus, the information derived from such an analysis is not likely to become obsolete. The nature and number of genetic changes, can provide clues to the history of the cancer cell. Finally, a high resolution genomic analysis may lead to the discovery of new genes involved in the etiology of the disease or disorder of interest.

Microarrays typically have many different DNA molecules, often referred to as probes, fixed at defined coordinates, or addresses, on a flat, usually glass, support. Each address contains either many copies of a single DNA

probe, or a mixture of different DNA probes, and each DNA molecule is usually 2000 nucleotides or less in length. The DNAs can be from many sources, including genomic DNA or cDNA, or can be synthesized oligonucleotides. For clarity and brevity, we refer to those chips with genomic or cDNA derived probes as DNA chips and those chips with synthesized oligonucleotide probes as oligo chips, respectively. Chips are typically hybridized to samples, applied as single stranded nucleic acids in solution.

Hybridization with samples at a given address is determined by many factors including the concentration of complementary sequences, the probe concentration, and the volume from which each address is able to capture complementary sequences by hybridization. We refer to this volume as the diffusion volume. Because the diffusion volume, and hence, the potential hybridization signal, may vary from address to address in the hybridization chamber, the probe array is most accurate as a comparator, measuring the ratio of hybridization between two differently labeled specimens that are thoroughly mixed and therefore share the same hybridization conditions, including the same diffusion volume. Typically the two specimens will be from diseased and disease free cells.

We distinguish between compound and simple DNA probe arrays based on the nucleotide complexity of the probes at each address. When this nucleotide complexity is less than or equal to about 1.2 kb per address, we speak of simple DNA probe arrays. When it exceeds 1.2 kb per address, we speak of compound probe arrays. Simple probe arrays are currently able to detect cDNA species that are present at 2 to 10 copies of mRNA per cell when contacted with a solution containing a total cDNA concentration of 1 mg/ml. The threshold of detection of a given species is estimated to be in the range of 4 to 20 ng/ml. Because a simple probe array is generally able to capture only a single species of DNA from the sample, this detection threshold poses a problem for the use of simple DNA probe arrays for analysis of

genomic DNA. The concentration of a unique 700 bp fragment of human genomic DNA (which has a total complexity of about 3000 mb) in a solution of total genomic DNA dissolved at its maximum concentration of 8 mg/ml would be about 2 ng/ml, just below the lower estimate of the threshold of detection. Hence, in its unaltered format, the simple DNA probe chip would not suffice for the robust detection of genomic sequences.

The compound chip partially addresses this problem by increasing the nucleotide complexity of different probes at a given address, allowing for the capture of several species of DNA fragments at a single address. The signals of the different captured species combine to yield a detectable level of hybridization from genomic DNA. Present forms of compound probe arrays place the insert found in a single clone of a megacloning vector, such as a BAC, at each address. Because each address contains fragments derived from the entire BAC clone, several problems are created. The presence of repeat elements in the genomic inserts requires quenching with cold unlabeled DNA. Also, the great size of the megacloning vector inserts limits the positional resolution. For example, in the case of a compound probe array made of BACs, hybridization to a particular address reveals only to which BAC the hybridizing sequence is complementary, and does not reveal the specific complementary gene or sequence within that BAC. Another drawback is the presence of DNA derived from the megacloning vector and host sequences. The steps of excising and purifying the genomic DNA inserts from the vector and host sequences complicate and hinder rapid fabrication of microarrays.

2.2. PROBLEMS ASSOCIATED WITH GENETIC ANALYSIS

Analysis of the genetic changes in human tumors is often problematic because of the presence of normal stroma. Samples of tumor tissue are often contaminated with non-cancerous cells, making isolation and study of tumor cell DNA difficult. While either microdissection or flow cytometry

can produce small samples highly enriched for tumor cells or nuclei, the amount of extracted DNA recoverable from such enriched samples is insufficient for most uses.

One technique which can be used on small samples is representational difference analysis (RDA). (U.S. Patent No. 5, 436, 142, Lisitsyn, et al., 1993, Science, 259:946-951) RDA is a subtractive DNA hybridization technique that is useful, e.g., to discover the differences between paired normal and tumor genomes. The first step of RDA requires making an "amplicon representation", which is a highly reproducible simplification and amplification of a DNA population. Typically, an amplicon representation is a set of restriction endonuclease fragments of a limited size range generated by PCR (polymerase chain reaction). PCR generates sufficient amounts of DNA for subsequent processing, on the order of 100 ug, starting from as little as 3 ng of DNA (the amount of DNA isolatable from about 1000 cells).

One limitation of the amplicon useful in RDA is that an amplicon representation with much lower complexity than that of the genome from which the amplicon is derived is needed to enable the subtractive hybridization to proceed effectively. Such low complexity representations (LCRs) do not "capture" enough (typically, 7% or less) of the genome to be generally useful for other applications. The complexity of the representation is related to the frequency of cutting of the restriction enzyme used to generate the genomic fragments, combined with the amplification reaction steps, e.g., PCR, which tend to favor the smaller fragments.

Whole genome amplification (WGA) is a method by which more complex amplifications of the DNA from minute samples are generated. (Sun, F, et al., 1995, Nucleic Acids Res. 23(15):3034-3040, Barrett, M.T., et al., 1995, Nucleic Acids Res. 23(17):3488-3492.) In WGA, PCR is performed on DNA isolated from small amounts of sample using random primers.

There are at least three disadvantages to the WGA method:

1. The amplified DNA can not be used for Southern analysis. Because more than one primer can bind to a single gene, a heterogenous mixture of different sized fragments can be generated from a single gene. This would result in a smear, not a band, being detected by Southern hybridization.

2. Due to the random nature of the amplification, each amplification results in a different mixture of fragments. Therefore the amplification is not reliably reproducible. This makes the use of such whole genomic amplifications for the purposes of sample to sample comparisons difficult.

3. Whole genomic amplifications are not useful for quantitating the copy number of genes present in the original sample. Because the primers are random, the representation of each gene can vary greatly with respect to the other genes. Thus, the abundance of each gene relative to other genes in the original sample is not preserved during the amplification, making quantitation of copy number impossible.

Thus, there continues a long felt need for a method of obtaining amounts of genetic material from scant genomic samples to enable genetic analysis of small samples using techniques which previously were inapplicable due to the limited amount of DNA isolatable from such samples. There is also a long felt need for a method of amplifying and storing DNA from scant, nonrenewable sources.

3. SUMMARY OF THE INVENTION

It is an object of the present invention to provide a solution to problems associated with the use of microarray technology for the analysis DNA. The present invention provides compositions and methods for the use of simple and compound representations of DNA in microarray technology. A representation of DNA is a sampling of DNA produced by a restriction endonuclease digestion of genomic or other DNA, followed by linkage of adaptors and then amplification with

primers complementary to the adaptors. The DNA may be from any source. Sources from which representations can be made include, but are not limited to, genomic or cDNA from tumor biopsy samples, including breast cancer and prostate cancer biopsies, normal tissue samples, tumor cell lines, normal cell lines, cells stored as fixed specimens, autopsy samples, forensic samples, paleo-DNA samples, microdissected tissue samples, isolated nuclei, and fractionated cell or tissue samples.

Representation of the genome results in a simplification of its complexity; the genomic complexity of a representation can range from below 1% to as high as 95% of the total genome. This simplification allows for desirable hybridization kinetics. Probes from representations of genomic DNA can be used as the probe of the microarray, and as the labeled sample hybridized to any microarray, however derived. Because formation of a representation involves the step of amplifying the DNA via an amplification reaction, such as the polymerase chain reaction, ligase chain reaction, etc., very small amounts of DNA can be used as starting material. The use of compound representations, defined as a representation of a representation, is also provided by the present invention. As is fully described below, compound representations can be used, for example, to screen for polymorphisms.

In addition, representational difference analysis (RDA), can be used for the efficient removal of vector and host sequences when constructing microarrays from megacloning vectors. RDA may also be used to remove any known, unwanted sequences from the representation, including repetitive sequences.

As used herein, the term "simple representation" refers to a sampling of DNA produced by a restriction endonuclease digestion of genomic or other DNA, followed by linkage of adaptors and then amplification with primers complementary to the adaptors.

As used herein, the term "compound representation" refers to a representation of a representation.

The present invention is also directed to methods for the production of High Complexity Representations (HCRs) of the DNA from cells. In one embodiment, the HCR is made by completely digesting a small amount of DNA from any source with a relatively frequent cutting restriction endonuclease, ligating adaptor oligonucleotides to the ends of the resulting fragments, and amplifying the fragments, for example by PCR, using primers to said adaptor oligonucleotides.

In another embodiment, the HCR is made by completely digesting a small amount of DNA from any source with at least two restriction endonucleases, ligating adaptor oligonucleotides to the ends of the resulting fragments, and amplifying the fragments, for example by PCR, using primers to said adaptor oligonucleotides.

HCRs can represent from 20% to 95% of the genome, depending on the restriction enzyme or enzymes used, and the conditions of the PCR amplification.

Sources from which HCR's can be made include, but are not limited to, tumor biopsy samples, including breast cancer and prostate cancer biopsies, normal tissue samples, tumor cell lines, normal cell lines, cells stored as fixed specimens, autopsy samples, forensic samples, paleo-DNA samples, microdissected tissue samples, isolated nuclei, and fractionated cell or tissue samples.

HCRs are useful for, but not limited to, determining gene copy number, deletion mapping, determining loss of heterozygosity, comparative genomic hybridization, and archiving of DNA.

4. BRIEF DESCRIPTION OF THE FIGURES

The present invention may be more fully understood by reference to the following detailed description of the invention, examples of specific embodiments of the invention and the appended figures in which:

Figure 1 (A-B) illustrates the results of PCR reactions designed to quantitate the complexity of HCRs. Panel A shows a gel on which the products of PCR reactions have been separated and visualized. The PCR reactions were performed using probes chosen randomly from an assortment of sequence tags, representing sequences known to be present in the human genome. This sequence tag is present in all of the HCRs produced from 14 tumor biopsy normals obtained by sorting (numbered 1-14). M represents the marker ϕx174 HaeIII digested. G denotes two different genomic DNAs used as positive controls, and - denotes a reaction which contained no DNA. Panel B shows the products of reactions performed on HCRs with 6 probes chosen randomly from an assortment of sequence tags, representing sequences known to be present in the human genome. The presence of these sequence tags in genomic DNA (G), HCR (H), and no DNA was tested. The HCR did not contain 2 of the 6 sequence tags assayed, as can be seen by the absence of a band in the HCR lane under the numbers 2 and 5. M represents the marker ϕx174 HaeIII.

Figure 2 (A-B) illustrates an analysis of copy number using low (LCR) and high (HCR) complexity representations and genomic DNA (Genomic) for several amplified loci (cycD1, c-erB2, and c-myc each denoting the respective locus). Panel a is a Southern blot comparing tumor cell lines (T) to normal (N). DpnII represents the HCRs and Bg1II represents the LCRs. The lane marked probe denotes the free probe used as a marker. The probes used for hybridization were derived from small Bg1II fragments isolated from P1 clones specific for each locus respectively. Figure 2B represents the quantitation of the above described Southern blots comparing the amount of amplification of high and low complexity representations with genomic DNA cut with the same restriction enzyme used to generate each representation.

Figure 3 illustrates the use of HCRs for deletion mapping. Shown is the deletion mapping of 7 tumor cell lines

(designated 1-7) which already display a known deletion pattern for several probes from the human genomic region 20p11. The deletion pattern of the DpnII HCRs (denoted HCR) is compared to the DpnII digest of the genomic DNA (denoted Genomic).

Figure 4 illustrates a comparison of primary tumor biopsies by HCR Southern blotting analysis. Primary tumor biopsy HCRs (denoted by a number preceded by BBR, CHTN, or NSBR) from matched diploid (Dpl) and aneuploid (Anu) were compared by Southern blot analysis. The c-myc probe which was hybridized was the same as that used in figure 2.

Figure 5 illustrates the results of a quantitative PCR analysis of HCRs. Diploid (Dpl) and aneuploid (Anu) HCRs derived from sorted primary tumor biopsies were used as template for QPCR analysis. Probes from several genomic regions (FHIT, p16, and c-erB2) were used to determine copy number in several HCRs. The data from the ABI 7700 Sequence Detector was analyzed with MS Excel to produce the graphs shown. The X axis represents the cycle number during the reaction and the Y axis denotes the fluorescence produced.

Figure 6 illustrates the use of HCRs for LOH analysis. Shown is LOH analysis carried out on HCRs derived from sorted primary tumor biopsies, where Dpl denotes diploid and Anu denotes aneuploid. The primers used in the reaction amplify a fragment from the p53 locus which contains a tetranucleotide repeat. +Gen denotes a mixed population normal genomic DNA which was used as positive control and +HCR denotes the HCR produced from this mixed normal genomic DNA. -lane represents a reaction which no template was added.

Figure 7 illustrates the use of HCRs for comparative genomic hybridization. Shown are two representative chromosome spreads (Ch 1, and Ch 17) comparing the genomic (Gen) to the HCR, for two different cell lines, BT474, and MCF7. Lines below the spreads denote differences which exceed the standard deviation, suggesting an abnormal copy number.

5. DETAILED DESCRIPTION OF THE INVENTION

The present invention provides for the use of simple and compound representations of DNA in microarray technology. Representations are used to obtain a reproducible sampling of the genome that has reduced complexity. A representational protocol initiates with restriction endonuclease cleavages followed by ligation of oligonucleotides to the cleaved DNA. Ultimately, these oligonucleotides are used for a gene amplification protocol such as PCR. The resulting representation can be advantageously applied to microarray technology as both the arrayed probe and hybridized sample.

Other protocols to produce a sampling of the genome of reduced complexity exist, but are not readily adaptable to microarray technology. The two most common methods are whole genome amplification (Telenius, et al, 1992, Genomics, 13, 718-25; Xu, K., et al, 1993, Hum Reprod, 8, 2206-10; Kristjansson, K., et al., 1994, Nat Genet, 6, 19-23; Sun, F., et al., 1995, Nucleic Acids Res, 23, 3034-40; Xiao, Y., et al., 1996, Cytogenet Cell Genet, 75, 57-62) and Inter ALU PCR (Cotter F.E., et al., 1991, Genomics, 9(3):473-80; Cotter F.E., et al., Genomics (1990), 7(2):257-63). Both of these methods have distinct disadvantages.

Whole genome amplification utilizes random primers for PCR amplification. Due to the use of random primers a random sampling of the genome results. More seriously, the random priming creates an enormous variability from PCR reaction to PCR reaction, due to the variable thermal stability of primers to template, resulting in variability in sampling that is hard to control or standardize.

Inter ALU PCR utilizes alu consensus primers to amplify the unique sequences between alu sequences. Only those fragments between alu sequences and small enough to be amplified by PCR are present in the Inter ALU samplings. Like the previously described method, the disadvantage of this method is that the sampling is highly dependent on PCR conditions, especially temperature. Due to the fact that the

primers for amplification are hybridizing to endogenous sequences, any mismatch between the primer and the recognition site in the alu sequence would cause a change in the representation. Any temperature fluctuation during amplification could create markedly different representations from the same sample if produced at different times. If this type of representation were used for microarray experiments, comparison from experiment to experiment would be difficult. These variations caused by inefficient amplification due to mismatch would make the production of a microarray based on this technique difficult if not impossible.

5.1. REPRESENTATIONS

It is an object of the present invention to provide for the use of representations of DNA in microarray technologies. Any such use is within the scope of the present invention and several non-limiting examples are described below.

A representation of DNA is a sampling of DNA, for example, the genome, produced by a restriction endonuclease digestion of genomic or other DNA, followed by linkage of adaptors and then amplification with primers complementary to the adaptors (Lucito, et al., 1998, Proc. Natl. Acad. Sci. USA, 95:4487-4492, incorporated herein by reference). Generally, only fragments in the size range of 200-1200 bp amplify well, so the representation is a subset of the genome.

Representations can be made from very small amounts of starting material (e.g., from 5ng of DNA), and are very reproducible. The reproducibility of representations has been demonstrated in several publications (Lisitsyn, et al., 1995, Proc. Natl. Acad. Sci. USA, 92:151; and Lucito, et al., 1998, Proc. Natl. Acad. Sci. USA, 95:4487-4492, both of which are incorporated herein by reference).

Any use of a simple or compound representation as a source for the probe attached to a chip, or as the sample hybridized to the chip, or as DNA from which a probe to be

hybridized to an array is derived, is within the scope of the invention. Arrays comprising probes derived from a representation by any method, for example by using the representation as a template for nucleic acid synthesis (e.g., nick translation, random primer reaction, transcription of RNA from represented DNA, oligonucleotide synthesis), or by manipulating the representation (e.g., size fractionation of the representation, gel purified fragments from the representation to the array) are also within the scope of the invention. Several applications of representations to DNA microarray technology are described below.

It is preferable that the one or more represented biological samples, and at least a fraction of the DNA comprising the microarray be from the same species. In a particular embodiment, the one or more samples are from a human, and at least a portion of the DNA on the microarray is human in origin. DNA from any species may be utilized according to the invention, including mammalian species (including but not limited to pig, mouse, rat, primate (e.g., human), dog and cat), species of fish, species of reptiles, species of plants and species of microorganisms.

5.2. REPRESENTATIONS HYBRIDIZED TO A NON-REPRESENTED SIMPLE OR COMPOUND PROBE ARRAY

In one embodiment, a representation of the DNA from one or more biological samples is hybridized to a microarray that is comprised of elements not from a representation. The microarray can be a simple or a compound array. In a particular embodiment, a representation of the DNA from one or more biological samples is hybridized to a compound probe array comprised of, for example, DNA from a megacloning vector such as a BAC, YAC, PAC, P1, or cosmid. In another embodiment, the DNA in the array derives from expressed sequences such as may be obtained from cDNAs or expression sequence tags (ESTs). The DNA in the array, in these embodiments, is not from a representation. In a particular

embodiment, the one or more samples hybridized to the microarray are from a human, and the microarray is comprised of DNA from one or more megacloning vectors that contain human DNA inserts. The represented samples may derive from any DNA, e.g., cDNAs or genomic DNAs, and may be high or low complexity representations. In a further embodiment, two represented samples are used, and the samples are differentially labeled so that hybridization of each sample can be individually quantitated and compared to the other sample. Differential labeling can be done with two different fluorescent indicators, e.g., Cy5-dCTP, fluorescein-dCTP, or lissamine-5-dCTP. Differential labeling and the hybridization of so labeled DNAs to microarrays are known to those of skill in the art (Schena, et al., 1995, Science, 270:467-470; Schena, et al., 1996, Proc. Natl. Acad. Sci. USA, 93:10614-19; Schena, et al., 1996, BioEssays, 18(5):427-31; Shalon, et al., 1996, Genome Research, 6:639-645, each of which is incorporated herein by reference).

This embodiment is useful for detecting variations in gene copy number. For example, representations of genomic DNA taken from a normal sample and genomic DNA taken from a sample of a tumor biopsy from a human can be differentially labeled and hybridized to a microarray fabricated with a BAC library spanning a significant portion of the human genome. Fixed at each address of the microarray is DNA from a single, different member of the BAC library. The hybridization signal from the tumor sample can be detected and compared to that of the normal sample. The signals at most of the addresses should be similar, but an address where, for example, the tumor sample has greater fluorescence, indicates that there has been an amplification in the tumor cell genome of the sequences corresponding to the BAC insert DNA at that address. This embodiment can also be useful for detecting variations in levels of gene expression when the represented sample derives from cDNA. This embodiment can also be used to assess the reproducibility of representations by comparing hybridization patterns of different representations from the

same sample. Similar or identical patterns of hybridization indicate that the representations are reproducible.

5.3. SAMPLES HYBRIDIZED TO A REPRESENTED COMPOUND PROBE ARRAY

In a second embodiment, both the compound probe array DNA and the hybridized sample DNA are from representations. The difference from the embodiment described in section 5.2 is that the microarray is fabricated with compound probes derived from a representation. A represented compound array has DNA sequences from more than one fragment of the representation at each address. The decreased complexity of both the array and sample DNAs allows for favorable hybridization kinetics and improved detection. Preferably, the sample and microarray DNA are identically represented, i.e., cut with the same restriction enzyme, ligated to the same adaptors, and amplified via, for example, the polymerase chain reaction. This embodiment can also be used to assess the reproducibility of representations by comparing hybridization patterns of different representations from the same sample. Similar or identical patterns of hybridization indicate that the representations are reproducible.

One technique which can be used whenever the DNA to be fixed to the microarray is from a megacloning vector is representational difference analysis (RDA). (U.S. Patent No. 5, 436, 142, Lisitsyn, et al., 1993, Science, 259:946-951) RDA is a subtractive DNA hybridization technique that is used to remove vector specific sequences, leaving substantially only the insert DNA. RDA can also be used to remove any other unwanted DNA sequences from the DNA to be fixed to the array, or the sample DNA. Such sequences can include repetitive DNA sequences.

In another embodiment, the compound probe array DNA is from a representation, and the hybridized sample DNA is any DNA, whether from a representation or not.

5.4. SAMPLES HYBRIDIZED TO A REPRESENTED SIMPLE PROBE ARRAY

In another embodiment, a simple probe array made from a representation is hybridized to a sample comprising DNA, whether from a representation or not.

In yet another embodiment, the sample DNA hybridized to the microarray is a representation of DNA, e.g., genomic DNA, and the microarray is a simple probe array fabricated with a representation of DNA. A represented simple probe array has DNA from only one fragment of a representation at each address. Thus, each element of the array comprises many copies of a single DNA molecule derived from a representation of genomic DNA.

The arrayed probes of any array may, if so desired, be mapped to any known library of genomic DNA. For example, the method of orthogonal partition hybridization can be used to map DNA libraries derived from genomic DNA or representations of genomic DNA to inserts of megacloning vector libraries. Libraries of probes from representations of the total genome, which can be used later for arraying, can be mapped. The probe library could be converted into 96 well dishes, and the collection maintained by PCR and manipulated robotically. The map positions of most of the probes can be determined after arraying, and records kept electronically. Those probes that cannot be mapped to the library of genomic DNA can later be mapped either as needed or as new mapping tools become available.

Arrays of simple DNA probes can be mapped, for example, by hybridization to orthogonal partitions of libraries of megacloning vectors. This can be illustrated by the following, non-limiting example: the assignment of arrayed probes to a positionally mapped megaYAC library of about 10,000 elements. Although this example is oriented towards YACs, because an ordered collection exists, the same principles can be applied to mapping arrays of simple probes to other ordered collections of vectors.

A partition is the division of a set into subsets, such that every element of the set is in one and only one

subset. Two partitions are called orthogonal if the intersection (i.e., the common elements) of any two subsets, one from each partition, contains no more than one element of the original set. If the members of the original set are arbitrarily laid out as a square, it is easy to see that there are always at least two mutually orthogonal, and in this case, equal partitions. These can be thought of as the partition of rows, and the partition of columns. There is a third mutually orthogonal and equal partition, the partition of "wrapped" diagonals, which will not be utilized in this example. These partitions have the additional property that each subset from one intersects each subset from the other in exactly one element. Each subset of one partition intersects a subset from another partition at a single element, and every element is the intersection of three subsets. Applying these ideas to a YAC library of 10,000 members, we see that it is possible to make two equal and orthogonal partitions of this library, each partition having about 100 subsets of about 100 members each. Many other pairs of orthogonal partitions can be envisioned, in particular, ones with larger numbers of smaller subsets.

Hybridization with representations of subsets from two orthogonal partitions could then be performed. If a probe hybridized to two subsets, one from each partition, that probe should have sequences in common with their intersection, which would be a unique YAC, if no YACs overlapped.

Because the YACs in a large library, such as the library contemplated in this example, will overlap, and a given probe may be in two or more members of the library, probes may hybridize to more than one subset of a partition. For example, if a probe is contained in two overlapping YACs, and hence hybridized to two subsets in each partition, there will be two possible solutions, with four candidate YACs, to the hybridization pattern with two orthogonal partitions. Knowledge of the mapping assignments of the YACs should be

sufficient to resolve this ambiguity. Only one pair of YACs will be neighbors.

The case for a probe contained in three overlapping YACs is only slightly more complex. There are more possible solutions to the hybridization pattern with two orthogonal partitions: 6 possible triads, picked from nine possible candidate YACs. Even these ambiguous cases can be resolved from a knowledge of the chromosomal assignment of the YACs. The odds that three YACs picked at random all derive from the same chromosome is roughly the square of the reciprocal of the number of chromosomes (1/23), or roughly 1 in 500. If there are six possible triads of YACs that may contain the probe, it is highly likely (nearly 99% odds) that only the true triad will derive from the same chromosome. The success of resolution increases, approaching completeness, when we consider the finer map assignments of the YACs. The chances that three YACs, picked at random, are all neighbors is vanishingly small.

This embodiment is useful for detection of changes in gene copy number between normal and, for example, cancer biopsy samples, as is described in section 5.2. If the elements have been mapped, as described above, positional information of the alteration of gene copy number can be gathered. This embodiment can also be used to assess the reproducibility of representations by comparing hybridization patterns of different representations from the same sample. Similar or identical patterns of hybridization indicate that the representations are reproducible.

5.5. COMPOUND REPRESENTATIONS HYBRIDIZED TO A REPRESENTED SIMPLE PROBE ARRAY

In another aspect of the invention, the sample DNA hybridized to the microarray is or is derived from a compound representation of DNA, and the microarray, like that described in section 5.4, is a represented simple probe array. A compound representation is the result of two or more consecutive representations. In its simplest form, a

compound representation is made by making a first representation of, for example, genomic DNA, followed by the making of a second representation of the first representation. Preferably, different restriction enzymes are used for each sample representation, and the enzyme used to prepare the first sample representation and the representation immobilized on the microarray are the same.

The following non-limiting example will serve to illustrate two possible compound representations termed AcB and AsB, where A and B are any two restriction endonucleases. They derive from a first representation made by using the A restriction endonuclease. This first representation will consist of fragments that have an A restriction endonuclease site at each end, such fragments are termed AA fragments. A fragment with a B restriction endonuclease site at each end is termed a BB fragment, while fragments with an A restriction endonuclease site at one end and a B restriction endonuclease site at the other is termed an AB fragment. AcB representations consist of AB and BB fragments that derive from those AA fragments of the simple A representation that contain a B restriction endonuclease site. AcB representations consist of those AA fragments of the simple A representation that do not contain a B restriction endonuclease site.

AsB is made by making a first representation with the restriction endonuclease A, then making a second representation by cleaving the resulting AA fragments with the restriction endonuclease B and amplifying with the same primers used in the first representation. AA fragments from the first representation that have an internal B site are cut by the B restriction endonuclease and will not amplify, while those AA fragments lacking an internal B site will amplify. The final representation then consists only of those AA fragments with no internal B site.

The second representation, designated AcB, is also made from a first simple A representation, i.e., a representation made with restriction endonuclease A. AcB is

made, like AsB, by making a first representation with the restriction endonuclease A, then making a second representation by cleaving the resulting AA fragments with the restriction endonuclease B. This cleavage results in three types of fragments: 1) AA fragments, i.e., those AA fragments without internal B sites, 2) AB fragments, i.e., fragments with an A site at one end and B site at the other, derived from those AA fragments with one or more internal B sites, and 3) BB fragments, derived from those AA fragments with more than one internal B site. The difference between the AcB and AsB representations is in the amplifications steps of the second representation. In the AcB representation, oligonucleotide adaptors ("B adaptors") are ligated at the B site on both the 5' and 3' ends. Then, an A adaptor is ligated to the 5' end only. This adaptor has a different sequence than the adaptor used for the first, simple representation, and is much longer, on the order of 40 nucleotides. After ligation, and removal of unligated adaptors, the ability of these molecules to extend from the 3' end is removed by dideoxy extension. Finally, primers to the A and B adaptors are added and the product is exponentially amplified by PCR using a polymerase without 3' exonuclease activity. Only AB and BB fragments are strongly favored to amplify.

The protocol for AcB may seem more complex than needed. The reason for adding the A adaptor to the 5' end only is to disable exponential amplification from strands that have A at both ends. Even with this step, there will be some AA fragments that reanneal during the polymerase chain reaction step, fill-in at their 3' ends during the chain elongation step, and subsequently amplify from the A oligonucleotide primer, thereby poisoning the representation. Hence, two more features are added. The new A adaptor is long (40 nucleotides or longer). Those AA molecules that do form, and become adapted at their 5' and 3' ends by self priming, will amplify very poorly because the length of the adaptor will create thermally stable "pan handles", as

described above. Finally, the 3' ends of all A sites are blocked by dideoxy extension to reduce the possibility of self priming after reannealing, and the subsequent formation of amplifiable AA fragments.

AcB and AsB representations are useful for detecting internal polymorphic restriction endonuclease sites and for detecting heterozygous and homozygous states with respect to those polymorphic sites. When a simple DNA probe chip made with a simple A representation (i.e., the first representation of the AcB or AsB compound representations) is hybridized with differentially labeled AcB (for this example, a red label) and AsB (for this example, a green label) representations, both homozygous and heterozygous states are readily detected: high red ratios indicate both alleles have B sites; high green ratios indicate both alleles do not have B sites; and ratios near equality (yellow) indicate the heterozygous state. In a preferred embodiment, the second restriction endonuclease, i.e., the B restriction endonuclease is one that recognizes CpG, such as TaqI. Such restriction endonucleases are especially useful since the sequence CpG is especially polymorphic.

5.6. PREPARATION OF REPRESENTATIONS

Briefly, representations are generated by restriction endonuclease digestion of DNA, followed by linkage of adaptors and then amplification with primers complementary to the adaptors. The DNA may be from any source. The method is adaptable to any genome. It is often advantageous to isolate DNA contemporaneously from both normal and diseased cells, for example, from normal and cancerous tissue, preferably from the same individual. Parallel processing of the samples allows for more accurate comparisons of the representations generated from the two different sources of cells.

The DNA is isolated by any convenient means, and then substantially completely digested by any means, such as

the use of a restriction enzyme endonuclease, which results in cutting at predetermined sequences.

The complexity of the representation can be shaped in several ways. High complexity representations (HCRs) are obtained by cleaving the DNA of interest with a relatively frequent cutting restriction enzyme, such as DpnII. This results in a majority of the fragments being between 200-1200 bp, and therefore amplifiable. Representations derived from DpnII digests have about 70% of the complexity of the entire genome, i.e., 70% of the genome is present in such a representation.

Low complexity representations (LCRs) are obtained by cleaving the DNA of interest with a relatively infrequent cutting restriction enzyme, such as BamHI or BglII, resulting in a minority of the fragments being between 200-1200 bp. Representations derived from BamHI or BglII digests have about 2%, the complexity of the entire genome.

A restriction enzyme which is inhibited by methylation of the DNA can be selected for the digestion step. The use of such an enzyme can reveal differences in methylation between compared samples. This can be useful because, for example, it has been suggested that there are differences in methylation between normal cells and some cancerous cells.

Complexity of the representation can also be shaped by the adaptors used for amplification. Because the same adaptors are used at both ends of the cleaved fragments, the single strands form panhandles (Lukyanov, G.A., et al., 1995, Anal. Biochem. 229:198-202, incorporated herein by reference). This inhibits amplification by PCR, because panhandle formation competes with PCR primer annealing, a necessary step for amplification. Shorter fragments are preferentially inhibited due to the close proximity of the adaptors resulting effectively in a higher local concentration of the 5' and 3' adaptors linked to the ends of such fragments, as compared with longer fragments. Adaptors that form panhandles of 29 nucleotides allow for

amplification of fragments in the size range of 200-1200 bp. Shorter adaptors that form panhandles of 24 nucleotides release some of the inhibition of the smaller fragments, resulting in the favoring of smaller PCR amplification products, and therefore, a representation of altered complexity.

The DNA may be from any source. Sources from which representations can be made include, but are not limited to, tumor biopsy samples, including breast cancer and prostate cancer biopsies, normal tissue samples, tumor cell lines, normal cell lines, cells stored as fixed specimens, autopsy samples, forensic samples, paleo-DNA samples, microdissected tissue samples, isolated nuclei, and fractionated cell or tissue samples.

The degree of complexity of the representation generated is related to the frequency of cutting, specifically, more frequent cutting enzymes will result in higher complexity representations. Thus, representations of the desired complexity can be produced by the selection of the appropriate enzyme. The selection can be made with the guidance of the art, including readily available information on the frequency of cutting of various enzymes and the average fragment lengths generated by said enzymes (Bishop, et al., 1983, A Model For Restriction Fragment Length Distributions, Am. J. Hum. Genet. 35:795-815). To prepare representations from highly degraded DNA it may be preferable to use restriction endonucleases that cleave with relatively greater frequency than, for example, a restriction enzyme such as DpnII.

After digestion of the DNA, the oligonucleotide adaptors are ligated to the ends of each of the strands of the DNA. The adaptor will usually be staggered at both ends, with one strand being longer than the other and therefore being single stranded over a small region at the end not ligated to the digested fragments. In the case when the restriction enzyme digestion leaves staggered ends, the

adaptor will have an end complementary to the fragments' staggered ends.

The DNA is then amplified by an amplification reaction, for example, by adding primer and using the polymerase chain reaction for usually at least 15 cycles and generally not more than about 35 cycles. The primer will be complementary to the adaptor. The adaptors are then removed by restriction endonuclease digestion and separation, using any convenient means.

For purposes of comparing representations from two different sources, it is preferable that HCRs are prepared from the same amount of starting material, that the genomic DNAs are extracted in the same manner, and that PCR is performed at the same time under the same conditions in the same thermal cycler.

5.7. PREPARATION OF MICROARRAYS

Microarrays for use in the present invention are known in the art and consist of a surface to which probes can be specifically hybridized or bound, preferably at a known position. Each probe preferably has a different nucleic acid sequence. The position of each probe on the solid surface is preferably known. In one embodiment, the microarray is a high density array, preferably having a density of greater than about 60 different probes per 1 cm^2.

To manufacture a microarray DNA probes are attached to a solid support, which may be made from glass, plastic (e.g., polypropylene, nylon), polyacrylamide, nitrocellulose, or other materials, and may be porous or nonporous. A preferred method for attaching the nucleic acids to a surface is by printing on glass plates, as is described generally by Schena et al., 1995, Science 270:467-470. See also DeRisi et al., 1996, Nature Genetics 14:457-460; Shalon et al., 1996, Genome Res. 6:639-645; and Schena et al., 1995, Proc. Natl. Acad. Sci. USA 93:10539-11286.

A second preferred method for making microarrays is by making high-density oligonucleotide arrays. Techniques are

known for producing arrays containing thousands of oligonucleotides complementary to defined sequences, at defined locations on a surface using photolithographic techniques for synthesis in situ (see, Fodor et al., 1991, Light-directed spatially addressable parallel chemical synthesis, Science 251:767-773; Pease et al., 1994, Light-directed oligonucleotide arrays for rapid DNA sequence analysis, Proc. Natl. Acad. Sci. USA 91:5022-5026; Lockhart et al., 1996, Expression monitoring by hybridization to high-density oligonucleotide arrays, Nature Biotech 14:1675; U.S. Patent Nos. 5,578,832; 5,556,752; and 5,510,270, each of which is incorporated by reference in its entirety for all purposes) or other methods for rapid synthesis and deposition of defined oligonucleotides (Blanchard et al., 1996, High-Density Oligonucleotide arrays, Biosensors & Bioelectronics 11: 687-90). When these methods are used, oligonucleotides (e.g., 20-mers) of known sequence are synthesized directly on a surface such as a derivatized glass slide.

Other methods for making microarrays, e.g., by masking (Maskos and Southern, 1992, Nuc. Acids Res. 20:1679-1684), may also be used. In principal, any type of array, for example, dot blots on a nylon hybridization membrane (see Sambrook et al., Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989, which is incorporated in its entirety for all purposes), could be used, although, as will be recognized by those of skill in the art, very small arrays will be preferred because hybridization volumes will be smaller. Presynthesized probes can be attached to solid phases by methods known in the art.

5.8. PREPARATION OF SAMPLE NUCLEOTIDES

Sample to be hybridized to microarrays can be labeled by any means known to one of skill in the art. The sample may be from any source, including a representation, cDNA, RNA or genomic DNA. In a particular embodiment, the sample is labeled with a fluorescent probe, by, for example,

random primer labeling or nick translation. When the sample is a representation, it may be labeled during the PCR step of making the representation by inclusion in the reaction of labeled nucleotides. The fluorescent label may be, for example, a lissamine-conjugated nucleotide or a fluorescein-conjugated nucleotide analog. Sample nucleotides are preferably concentrated after labeling by ultrafiltration.

In a particular embodiment, two differentially labeled samples (e.g., one labeled with lissamine, the other fluorescein) are used.

5.9. HYBRIDIZATION TO MICROARRAYS

Hybridization of a representation of a sample to an array encompasses hybridization of the representation, or nucleotides derived from the representation by any method, for example by using the representation as a template for nucleic acid synthesis (e.g., nick translation, random primer reaction, transcription of RNA from represented DNA), or by manipulating the representation (e.g., size fractionation of the representation, gel purified fragments from the representation to the array).

Nucleic acid hybridization and wash conditions are chosen such that the sample DNA specifically binds or specifically hybridizes to its complementary DNA of the array, preferably to a specific array site, wherein its complementary DNA is located, i.e., the sample DNA hybridizes, duplexes or binds to a sequence array site with a complementary DNA probe sequence but does not substantially hybridize to a site with a non-complementary DNA sequence. As used herein, one polynucleotide sequence is considered complementary to another when, if the shorter of the polynucleotides is less than or equal to 25 bases, there are no mismatches using standard base-pairing rules or, if the shorter of the polynucleotides is longer than 25 bases, there is no more than a 5% mismatch. Preferably, the polynucleotides are perfectly complementary (no mismatches). It can easily be demonstrated that specific hybridization

conditions result in specific hybridization by carrying out a hybridization assay including negative controls (see, e.g., Shalon et al., supra, and Chee et al., 1996, Science 274:610-614).

Arrays containing double-stranded probe DNA situated thereon are preferably subjected to denaturing conditions to render the DNA single-stranded prior to contacting with the sample DNA. Arrays containing single-stranded probe DNA (e.g., synthetic oligodeoxyribonucleic acids) need not be denatured prior to contacting with the sample DNA.

Optimal hybridization conditions will depend on the length (e.g., oligomer versus polynucleotide greater than 200 bases) and type (e.g., RNA, DNA) of probe and sample nucleic acids. General parameters for specific (i.e., stringent) hybridization conditions for nucleic acids are described in Sambrook et al., supra, and in Ausubel et al., 1987, Current Protocols in Molecular Biology, Greene Publishing and Wiley-Interscience, New York. When the cDNA microarrays of Schena et al. are used, typical hybridization conditions are hybridization in 5 X SSC plus 0.2% SDS at 65°C for 4 hours followed by washes at 25°C in low stringency wash buffer (1 X SSC plus 0.2% SDS) followed by 10 minutes at 25°C in high stringency wash buffer (0.1 X SSC plus 0.2% SDS) (Shena et al., 1996, Proc. Natl. Acad. Sci. USA, 93:10614). Useful hybridization conditions are also provided in, e.g., Tijessen, 1993, Hybridization With Nucleic Acid Probes, Elsevier Science Publishers B.V. and Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press San Diego, CA.

5.10. DETECTION OF HYBRIDIZATION

Hybridization to the array may be detected by any method known to those of skill in the art. In a particular embodiment, the hybridization of flourescently labeled sample nucleotides is detected by laser scanner. When two different fluorescent labels are used, the scanner is preferably one

that is able to detect fluorescence of more than one wavelength, the wavelengths corresponding to that of each fluorescent label, preferably simultaneously or nearly simultaneously.

5.11. GENERATION OF HIGH COMPLEXITY REPRESENTATIONS

Briefly, High Complexity Representations (HCRs), are generated by restriction endonuclease digestion of DNA, followed by linkage of adaptors and then amplification with primers complementary to the adaptors. The DNA may be from any source. The method is adaptable to any genome. It is often advantageous to isolate DNA contemporaneously from both normal and diseased cells, for example, from normal and cancerous tissue, preferably from the same individual. Parallel processing of the samples allows for more accurate comparisons of the two HCRs generated from the two different sources of cells.

The DNA is isolated by any convenient means, and then substantially completely digested by any means, such as the use of a restriction enzyme endonuclease, which results in frequent cutting at predetermined sequences.

In one embodiment, a "relatively frequent cutting restriction endonuclease" is used. As used herein, the term "relatively frequent cutting restriction endonuclease" is intended to mean a restriction endonuclease which has a consensus sequence of four or fewer nucleotides, and may provide for blunt ends or staggered ends. Exemplary "relatively frequent cutting restriction endonucleases" include, but are not limited to DpnII, Tsp509I, MboI, Sau3A1, MaeII, MspI, HpaII, BfaI, HinPI, Csp61, TaqI, MseI, AluI, BstUI, DpnI, HaeIII, RsaI, HnaI, and NlaIII.

In an alternative embodiment, at least two restriction enzymes are used simultaneously or sequentially to cut DNA with the desired frequency. The enzyme combination used should be chosen such that at least about 50% of the fragments produced by the digestion will be between 100 and 1000 nucleotices in length. It is within the

skill in the art to select such combinations. (Bishop, et al., 1983, Am. J. Hum. Genet. 35:795-815, incorporated by reference herein).

A restriction enzyme which is inhibited by methylation of the DNA can be selected for the digestion step. The use of such an enzyme can reveal differences in methylation between compared samples. This can be useful because, for example, it has been suggested that there are differences in methylation between normal cells and some cancerous cells.

As detailed above, the degree of complexity of a representation is related to the frequency of cutting, specifically, more frequent cutting enzymes will result in higher complexity representations. Thus, representations of the desired complexity can be produced by the selection of the appropriate enzyme. To prepare HCRs from highly degraded DNA it may be preferable to use restriction endonucleases that cleave with relatively greater frequency than, for example, a restriction enzyme such as DpnII.

After digestion of the DNA, the oligonucleotide adaptors are ligated to the ends of each of the strands of the DNA. The adaptor will usually be staggered at both ends, with one strand being longer than the other and therefore being single stranded over a small region at the end not ligated to the digested fragments. In the case when the restriction enzyme digestion leaves staggered ends, the adaptor will have an end complementary to the fragments' staggered ends.

The DNA is then amplified by an amplification reaction, for example, by adding primer and using the polymerase chain reaction for usually at least 15 cycles and generally not more than about 35 cycles. The primer will be complementary to the adaptor. The adaptors are then removed by restriction endonuclease digestion and separation, using any convenient means.

For purposes of comparing HCRs from two different sources, it is preferable that HCRs are prepared from the

same amount of starting material, that the genomic DNAs are extracted in the same manner, and that PCR is performed at the same time under the same conditions in the same thermal cycler.

5.12. USES OF HIGH COMPLEXITY REPRESENTATIONS

HCR's are useful for, among other things, determining gene copy number, deletion mapping, loss-of-heterozygosity (LOH) and comparative genomic hybridization (CGH). HCR's are also useful for microarrays, as described above. HCRs also are a generally useful means of "immortalizing" and archiving DNA for later analysis.

5.12.1. ARCHIVING OF DNA

HCRs of from the DNA of nonrenewable sources can be produced and stored, creating an archivable representation of the DNA from the original source. Further analysis can then be performed on the HCR instead of on the limited amount of original material.

5.12.2. HCRs FROM STORED SAMPLES

HCRs can be prepared from normal and tumor tissue stored as fixed, paraffin embedded, archived biopsies, and this would greatly extend the utility of such samples. As compared with fresh samples, more rounds of PCR are usually required to obtain workable amounts of DNA. The amplified DNA from stored samples usually has a lower size distribution than HCRs prepared from DNA extracted from fresh sources. HCRs prepared from paired stored samples are similar to each other, which suggest that the method has utility.

5.12.3. DETERMINATION OF LOSS OR AMPLIFICATION OF GENES BY MEASUREMENT OF GENE COPY NUMBER

Genomes often contain either extra copies of sequences due to gene amplification, or missing sequence when genes are deleted, which is known as loss of heterozygosity when one allele of a gene is lost, or loss of homozygosity

when both alleles are lost. Comparison of Southern blots of HCRs from diseased cells and normal cells can reveal whether the gene corresponding to the probe (for example a probe for a tumor suppressor of oncogene) is amplified or missing in the diseased cells relative to normal cells.

Some variability in the content of HCRs will arise due to polymorphism. For example, if a given sequence in an individual is contained on a bi-allelic DpnII fragment, occurring on a large and small fragment, and the small fragment is lost in the tumor due to loss of heterozygosity, the HCR from tumor may appear to be missing the sequence in question because 1) the large fragment will not be efficiently amplified by PCR, and will not be represented in the library, and 2) the small fragment is not present in the starting material, due to LOH. This can be used in a rapid method for loss of heterozygosity analysis if a sufficient number of such polymorphic sequences were known.

5.12.4. COMPARATIVE GENOMIC HYBRIDIZATION WITH HCRs

Comparative genome hybridization (CGH) is a powerful tool for analyzing the global genomic changes in tumors. (Thomspon, C.T. and Gray, J.W., 1993, J. Cell. Biochem. Suppl. 17G:139-143, Kallioniemi, O.P., et al., 1993, Semin. Cancer Biol. 4(1):41-46) In CGH, DNA from a test sample is labeled and mixed with normal DNA that is labeled with a different fluorophore. This probe mixture is hybridized to a normal metaphase spread or other reference standard. Since the entire fluorescently labeled genome of the test sample is used to stain normal metaphase chromosomes, the intensity of the fluorescence at each location along the normal chromosomes is proportional to the copy number of gene sequences that bind there. The resulting fluorescence ratios of hybridized test DNA to normal DNA is measured. One can observe the gains and losses of whole chromosomes or insertions and deletions on a specific chromosome.

CGH can be performed with HCRs by flourescently labeling the HCR and using it in a CGH protocol.

6. EXAMPLES

6.1. GENERATION OF REPRESENTATIONS

Materials

Restriction endonucleases as well as T4 ligase were supplied by New England Biolabs, Inc. Ampli Tag was supplied by Perkin Elmer Inc. Oligo adaptors RBg124 (5'_AGCACT CTCCAGCCTCTCACCGCA-3') and RBg112 (5'-GATCTGCGGTGA-3') were synthesized by BioSynthesis. dNTPs were supplied by Pharmacia. Cell lines used were obtained through ATCC, grown in culture and DNA from them isolated.

Production of Representations

5 - 10 ng of genomic DNA was digested by the desired restriction endonuclease (DpnII to produce the High Complexity Representation and BglII to produce the Low Complexity Representation) as suggested by the supplier. The digest was purified by phenol extracting and precipitation. The digested DNA was ligated to adaptors RBg124 and RBg112. The ligating mixture contained the digested genomic DNA, 1X reaction buffer (from the supplier), 444pmoles of each adaptor and water to bring the volume to 30 ul. The reaction was placed at 55°C and the temperature slowly decreased to 15°C. After the reaction mixture reached 15°C, 400 units of T4 DNA ligase was added,, and the reaction mixture was incubated at 15°C for 12-18 hrs. The ligated material was split into two PCR tubes and amplified by PCR. The PCR reaction contained the ligated material, 1X PCR buffer (335 MM Tris-HCL,, pH8.8, 20mM $MgCl_2$, 80 mM$(NH4)_2SO_4$, 50 mM beta-mercaptoethanol, 0.5 mg/ml of bovine serum albumin), 0.32 mM dNTP's, 0.6 mM RBg124 adaptor, which was then overlaid with mineral oil. The reaction was placed in a thermal cycler preheated to 72°C and then 15 units of AmpliTaq was added to the tube. The thermal cycler was set to continue 72°C for 5 minutes, and then repeat 20 cycles of 1 minute at 95°C, and 3

minutes at 72°C. This was followed by an additional 10 minutes at 72°C. The reaction was purified by phenol-chloroform, and then precipitation.

6.2. ANALYSIS OF THE COMPLEXITY OF AN HCR

We first tested the reproducibility and complexity of DpnII HCRs. We analyzed 14 different HCRs, each made from 5ng of DNA prepared from diploid nuclei separated from tumor biopsies by flow cytometry, and each amplified for 25 rounds. In our first sampling, we designed pairs of PCR primers to detect sequence tag sites (STSs). STSs are sequences known to be present in the genome, to which no particular function has been attributed. We picked STSs that were not cleaved by DpnII, and used primer pairs that amplified a single band from total genomic DNA controls. Of these, 18 of 25 pairs (72%) were able to amplify the same molecular weight fragment from each HCR, and 7 generally failed to amplify from any HCR (see Figure 1 for representative data). Our results suggests that DpnII HCRs reproducibly contain the same elements and about 70% of the genome.

We performed a similar test with primer pairs derived from the locus encoding the PTEN tumor suppressor gene, for which locus we had the complete nucleotide sequence. In this way we were able to use primers derived from DpnII fragments of known size. DpnII fragments were chosen at random, and PCR primer pairs designed for each. 22 pairs amplified single fragments by PCR from control genomic DNAs. These pairs were tested against the panel of 14 HCRs. Table 1 lists the size of the DpnII fragments from which the probes were derived.

Table 1

Fragment	Fragment Size	Present (+), or Absent (-)
1	94	+
2	97	-
3	126	+
4	134	+
5	160	+
6	193	+
7	415	+
8	424	+
9	460	+
10	466	+
11	466	+
12	495	+
13	496	+
14	507	+
15	528	+
16	531	+
17	1040	+
18	1049	+
19	1246	+
20	1436	+
21	1760	+
22	3916	-

PTEN fragments in the HCR. Fragments have been arranged by size order starting with the smallest, in the column labeled fragment size. The fragments have been numbered as a reference in the column labeled fragment. The column labeled pres or abs details the results, whether a fragment is in the HCR for each fragment from the PTEN region. The symbol + denotes presence and - denotes absence.

20 pairs amplified the expected fragment from all HCRs, and 2 pairs failed to amplify from any. The fragments that were not in the HCR were the largest, 3916 bp, and one

of the smallest, 97 bp. 16,039 bp was included in the HCRs and 4013 bp were excluded. Thus, assuming our pick of fragments was random, the HCRs contained about 75% of this region.

If DpnII cleavage is nearly complete during the preparation of an HCR, we expect that no PCR primer pairs should readily amplify from an HCR when the amplified sequence has a DpnII site. To test this, we chose 5 primer pairs from the PTEN locus that amplified a single fragment containing a single internal DpII site. All 5 pairs amplified fragments from genomic DNA controls, and none amplified detectable fragments from the 14 HCRs.

We conclude that HCRs prepared in parallel from samples processed in a similar manner are reasonably reproducible and represent about 70% of the human genome.

6.3. MEASURING GENE COPY NUMBER IN HCRs

Tumor genomes often contain either extra copies of sequences due to gene amplification, or missing sequence when genes are deleted. To explore the utility of representations for measuring gene copy number, we first compared Southern blots of genomic DNA to blots of HCRs and LCRs. For this purpose, we prepared genomic DNA from tumor cell lines amplified at cyclin D1 (MDA-MB-415), or c-erB2 (BT474), or c-myc (SKBr3), or human placenta. HCRs and LCRs were made from cell line or placenta DNAs using DpnII or Bg1II, respectively. As probes we used small Bg1II fragments that we cloned from P1's containing inserts from the designated loci. The blots, shown in Figure 2, panel A, were quantitated by phosphourimaging. To normalize for loading differences, the blots were stripped and rehybridized with a single copy sequence probe. The normalized ratios of signal from tumor and normal are tabulated in Figure 2, panel B. The same relative copy number (tumor to normal) was determined from blots of representations as was determined from blots of genomic DNAs. This indicates that there was no significant distortion in relative copy number for these

probes during the preparation of either high complexity or low complexity representations when these are prepared in parallel from similar starting materials. That is, the ratio of gene "X" in tumor to a normal standard to the ratio of gene "Y" in tumor to the same normal standard is constant for genomic, LCR and HCR DNAs.

To explore the utility of HCRs for deletion mapping, we probed blots of both genomic and HCR DNAs from tumor cell lines for deletion at the 20p11 locus. This locus was discovered initially using RDA, and subsequently found to be frequently deleted in gastrointestinal cancers. Figure 3 illustrates that the probe hybridized to sequences in the HCRs when and only when it hybridized to sequences in the respective genomic DNA.

6.4. PRODUCTION AND USE OF HCRs FROM SMALL SAMPLES

We tested the value of the HCRs made from limiting amounts of DNA. HCRs were prepared from aneuploid and diploid nuclei sorted from several breast cancer biopsies, and blotted for c-myc. Figure 4, illustrates that c-myc is amplified in the HCRs made from the aneuploid nuclei of some biopsy samples. We obtained confirmation of the validity of the c-myc amplifications by demonstrating that probes adjacent to but distinct from the c-myc probes were also amplified in the same samples.

6.5. MEASUREMENT OF GENE COPY NUMBER IN HCRs BY QUANTITATIVE PCR

Samples were tested by quantitative PCR. For this purpose, the fluorescence energy resonance transfer hybridization probes and the ABI7700 sequence detector were used to compare pairs of HCR DNAs prepared from aneuploid and diploid nuclei. The results, shown in Figure 5, indicate that no differences in copy number were detected by probes from an uninvolved region on chromosome 3. The curve for the aneuploid HCR amplified for c-erB2 arises 4 cycles sooner than does the curve for the paired diploid HCR, indicating a

higher copy number for c-erB2 of about 2exp4(16) fold in this sample. The curve for the aneuploid HCR deleted for p16 arises 4 cycles later than the paired diploid HCR, again a 16 fold difference, probably reflecting about 6% contamination of the aneuploid nuclei with diploid nuclei. One tumor/normal pair showed a shift of a single cycle for primer pairs detecting the p16 gene. This might reflect loss of a single allele in the tumor.

6.6. DETECTION OF LOSS OF HETEROZYGOSITY IN HCRs

Loss of heterozygosity (LOH) is a common lesion found in cancer cells, and may be indicative of genomic instability or the loss of function of a specific tumor suppressor gene. The detection of LOH is often obscured by the presence of normal stroma, and hence we tested whether HCRs prepared from minute amounts of samples highly enriched for tumor nuclei could be used for LOH analysis. PCR primers that amplify microsatellites and detect fragment length polymorphisms are frequently used for LOH mapping, and we chose to examine a primer pair that amplifies a highly polymorphic tetranucleotide repeat near the p53 locus.

Preliminary experiments established that these PCR primers detected the same allele pattern in both genomic and HCR DNAs prepared from cell lines. Twelve pairs of HCRs prepared from aneuploid and diploid nuclei were next examined. LOH at this locus was clearly detected in 9 out of 10 informative pairs (see Figure 6 for representative cases). This is greater than the reported proportion of LOH at this locus in breast cancer (60%), but may be biased for the highly aneuploid tumors which are sortable.

6.7. COMPARATIVE GENOME HYBRIDIZATION WITH HCRs

Comparative genome hybridization (CGH) is a powerful tool for analyzing the global genomic changes in tumors. (Thomspon, C.T. and Gray, J.W., 1993, J. Cell. Biochem. Suppl. 17G:139-143, Kallioniemi, O.P., et al., 1993, Semin. Cancer Biol. 4(1):41-46). The applicability of HCRs

to CGH was examined. For this experiment, tumor cell lines were chosen so that direct comparison of CGH performed with genomic and HCR DNA was possible. Little difference in patterns could be discerned with the two cell lines examined, BT474 and MCF7. Figure 7 shows the chromosomal scanning profiles obtained for two representatives chromosomes with each DNA source.

6.8. PREPARATION OF A MICROARRAY

In a specific exemplary embodiment, preparation of a microarray involves the steps of preparing the glass surface, preparing probes, and depositing the probes on the surface. Exemplary protocols for these steps are presented in this subsection.

Preparation of Poly-l-lysine Slides

Use 30 slide racks in 350mL glass dishes.

1. Dissolve 50g of NaOH pellets into 150ml ddH_2O
2. Add 200ml of 95% EtOH, stir until completely mixed
3. If solution remains cloudy, add ddH2O until clear
4. Pour solution into glass slide box.
5. Drop in 30 slides in a metal rack. (Gold Seal slides, Cat. 3010)
6. Let soak on an orbital shaker for at least two hours
7. Rinse slides by transferring rack to slide dish filled with ddH_2O
8. Repeat ddH_2O rinses x3. It's important to remove all traces of the NaOH-ethanol.
9. Prepare Poly-l-lysine solution: Use Sigma Poly-l-lysine solution. Cat. No. 8920
10. Add 70mL poly-l-lysine to 280ml of water
11. Transfer slides to lysine solution and let soak for 1 hour.
12. Remove excess liquid from slides by spinning the rack of slides on microtiter plate carriers at 500rpm.
13. Dry slides at 40 degrees C for 5 minutes in a vacuum oven.

14. Store slides in a closed box for at least two weeks prior to use.
15. Before printing arrays, check a sample slide to make sure it's hydrophobic (water should bead off it) but the lysine coating is not turning opaque.

Arraying

1. Transfer PCR reactions to 96-well V-bottom tissue culture plates (Costar). Add 1/10 vol. 3M sodium acetate (pH 5.2) and equal volume isopropanol. Store at -20 C for a few hours.
2. Centrifuge in Sorvall at 3500 RPM for 45 min. Rinse with 70% EtOH, centrifuge again and dry.
3. Resuspend DNA in 12ul 3X SSC for a few hours and transfer to flexible U-bottom printing plates.
4. Spot DNA onto poly-l-lysine slides with an arrayer.

Post processing

1. Rehydrate arrays by suspending slides over a dish of warm double distilled water. (~1 minute)
2. Snap-dry each array (DNA side up) on a 100C hot plate for 3 seconds.
3. UV X-link DNA to the glass by using a Stratalinker set for 60 milliJoules.
4. Dissolve 5g of succinic anhydride (Aldrich) in 315mL of n-methyl-pyrrolidinone.
5. To this, Add 35mL of 0.2M NaBorate pH 8.0 (made by dissolving boric acid in water and adjusting the pH with NaOH), and stir until dissolved.
6. Soak arrays in this solution for 15 minutes with shaking.
7. Transfer arrays to 95C water bath for 2 minutes
8. Quickly transfer arrays to 95% EtOH for 1 minute.
9. Remove excess liquid from slides by spinning the lack of slides on microtiter plate carriers at 500rpm.
10. Arrays can be used immediately

6.9. LABELING OF SAMPLE AND HYBRIDIZATION TO A MICROARRAY

2.5 μg each of two samples to be hybridized to a microarray are random primer labeled using Klenow polymerase (Amersham), one with a lissamine-conjugated nucleotide analog (DuPone NEN) and the other with a fluorescein-conjugated nucleotide analog (BMB). The two labeled samples are combined and concentrated for hybridization using an ultrafiltration device (Amicon).

The 5 μg of combined sample DNA is concentrated to 7.5 μl of TE buffer, denatured in boiling water and snap-cooled on ice. Concentrated hybridization solution is added to a final concentration of 5 x SSC/.01% SDS. The entire 10 μl of labeled sample DNA is transferred to the microarray surface, covered with a coverslip, placed in a humidity chamber and incubated in a 60 C water bath for 12 hours. The humidity is kept at 100% by the addition of 2μl of water in a corner of the chamber. The slide is then rinsed in 5XSSC/.1% SDS for 5 minutes and then in .2 xSSC/.1% SDS for 5 minutes. All rinses are at room temperature. The array is dried, and a drop of antifade (Molecular Probes) applied to the array under a coverslip.

6.10. DETECTION OF HYBRIDIZATION

A laser scanner is used to detect the two-color fluorescence hybridization signals from 1.8-cm x 1.8-cm arrays at 20-μm resolution. The glass substrate slide is mounted on a computer-controlled, two-axis translation stage (PM-500, Newport, Irvine, CA) that scans the array over an upward-facing microscope objective (20X, 0.75NA Fluor, Nikon, Melville, NY) in a bi-directional raster pattern. A water-cooled Argon/Krypton laser (Innova 70 Spectrum, Coherent, Palo Alto, CA), operated in multiline mode, allows for simultaneous specimen illumination at 488.0 nm and 568.2 nm. These two lines are isolated by a 488/568 dual-band excitation filter (Chroma Technology, Brattleboro, VT). An epifluorescence configuration with a dual-band 488/568 primary beam splitter (Chroma) excited both fluorophores

simultaneously and directed fluorescence emissions toward the tow-channel detector. Emissions are split by a secondary dichroic mirror with a 565 transition wavelength onto two multialkali cathode photomultiplier tubes (PMT; R928, Hamamatsu, Bridgewater, NJ), one with an HQ535/50 bandpass barrier filter and the other with a D630/60 band-pass barrier filter (Chroma). Preamplified PMT signals are read into a personal computer using a 12-bit analog-to-digital conversion board (RTI-834, Analog Devices, Norwood, MA), displayed in a graphics window, and stored to disk for further rendering and analysis. The back aperture of the 20X objective is deliberately underfilled by the illuminating laser beam to produce a large-diameter illuminating spot at the specimen (5-μm to 10-μm half-width). Stage scanning velocity is 100 mm/sec, and PMT signals are digitized at 100 μsec intervals. Two successive readings are summed for each pixel, such that pixel spacing in the final image is 20 μm. Beam power at the specimen is ~5 mW for each of the two lines.

The scanned image is despeckled using a graphics program (Hijaak Graphics Suite) and then analyzed using a custom image gridding program that creates a spreadsheet of the average red and green hybridization intensities for each spot. The red and green hybridization intensities are corrected for optical cross talk between the fluorescein and lissamine channels, using experimentally determined coefficients.

6.11. RANDOM PROBES, WITH REPRESENTED SAMPLES OF HUMAN GENOMIC DNA

In the experiments of this example arrays are made with random probes (with an average length of 1 kbp) taken from the human genome. For this example we assume that a chip of 100,000 elements can be made. We choose this number for illustrative purposes.

These chips are hybridized with DNA derived from two human samples that are prepared from a tumor and from the normal cells of the same patient.

The arrays are hybridized under the conditions described in the literature (Schena, et al., 1995, Science, 270, 467-70; Schena, et al., 1996, Proc Natl Acad Sci U S A, 93, 10614-9; Schena, M., 1996, Bioessays, 18, 427-31; Shalon, D., et al., 1996, Genome Res, 6, 639-45). Variations in these conditions can be tested to optimize the ratio of signal to noise in the hybridization, as will be discussed in detail below.

In the first experiment, total human genomic DNA is used as samples. One sample (tumor) is labeled with a fluorescent dye having a certain emission wavelength and the other labeled with a fluorescent dye having a distinguishable emission spectra. We speak of these dyes as reading in the "green" and "red" channels, respectively. The labeling follows procedures that are available in the literature (Schena, et al., 1995, Science, 270, 467-70; Schena, et al., 1996, Proc Natl Acad Sci U S A, 93, 10614-9; Schena, M., 1996, Bioessays, 18, 427-31; Shalon, D., et al., 1996, Genome Res, 6, 639-45).

Most probes of the chip (about two thirds) do not display significant hybridization signal in either the red or green channels. This is because the complexity of the human genome is so great that very little hybridization to the single copy probes on the chip occurs. (About two thirds of the human genome are single copy sequences.) About one third of probes light up very brightly in both channels. On identifying and sequencing a sample of these bright "yellow" probes (high green and red fluorescence), they turn out to contain repetitive sequences. (About one third of the human genome is repetitive sequences.) The strong green and red signals are due to the abundance of repetitive sequences in the genome, and hence, unlike single copy DNA, hybridization is readily observable. No significant alterations in green-red ratios are observed for any probe, and the experiment is uninformative.

In the second experiment, Bgl II representations of the genomic DNA are used, and prepared as previously

described (Lisitsyn, et al., 1993, Science, 259, 946-51). Bgl II representations have a complexity of about 2% the complexity of the entire human genome.

The representations are labeled as before with distinguishable fluorescent dyes, "green" for tumor DNA from the biopsy and "red" with normal DNA from the same patient. The same arrays are hybridized with the two labeled representations.

The hybridized chips are then analysed by scanning in the red and green channels to derive information about the relative gene copy concentrations in tumor and normal DNAs. Most probes of the chip (about two thirds) do not display significant hybridization signal in either the red or green channels. We call these class A probes. Most probes are in this category because most probes are not repetitive nor share sequences with the BglII representations. Therefore only background flourescence is observed.

About one third of probes light up very brightly in both channels. We call these class B probes. On identifying and sequencing a sampling of these bright "yellow" probes (high green and red fluorescence), they turn out to contain repetitive sequences. The strong green and red signal is due to the abundance of repetitive sequences in both the genome and the BglII representation of the genome.

A smaller number of probes, perhaps about 2%, have nearly equal and measurably higher than background fluorescence in both the green and red channels. We call these class C probes. (The distinction between class B and class C probes is made more clear in the following example.) Upon sequencing a sampling of these probes, we find that almost all contain at least one BglII site, and many contain two. They are showing detectable hybridization because they share sequences in common to BglII representations. There will be a total of about 2,000 of such probes (2% of 100,000).

An even fewer number of probes, perhaps only 0.1% of 2%, or a total of about 2, display significantly stronger

hybridization in the green channel than in the red channel. These class D probes, upon retrospective hybridization analysis of tumor and normal DNA by Southern blotting, are found to be significantly amplified in the tumor, indicating specific genetic lesions within the tumor. The estimate of 0.1% of 2% is based on the following. The amount of the genome that becomes amplified in an average tumor is about 3 megabases, or 0.1% of the genome, and the total number of probes in the array that share sequences with the represented sample of the genome is about 2%, as already stated.

These experiments illustrate two major points. The first point is that it is advantageous to reduce the nucleotide complexity of the sample to observe hybridization signal from single copy genomic sequences. According to the present invention in this example, we achieve this by making representations of the sample. The degree to which this complexity may be reduced is in part a function of the hybridization conditions and background noise, but reductions on the order of minimally ten fold and optimally about fifty fold are advantageous. The second point is that when using a represented sample, most randomly chosen probes are not very informative. Only those that share sequences with the represented sample are informative, and these are in a great minority. This can be remedied as is illustrated in the next examples.

These two major points are not necessarily essential when analysing lower complexity DNA populations, such as cDNA collections or genomic DNA from simpler organisms such as microbes, insects and some plants, wherein hybridization reactions go more to completion. For analysing cDNA populations, there are other reasons for preferring to represent samples or probes or both, as will be discussed in Example 6.19.

6.12. CULLING RANDOM PROBES, LEADING TO MORE INFORMATIVE ARRAYS

In the above example, most probes from the array were uninformative, either because they were unable to be used to detect hybridization to single copy DNA from the represented sample (class A) or because they recognized repeat sequences (class B). Only about 2000 probes (class C and D) were truly informative, indicating that no amplification (class C) or some amplification (class D) had occurred in the tumor.

It is possible to distinguish probes of class C or D from those of A or B, and to "cull" these probes to assemble a new array that is more efficient at detecting genetic differences between represented samples. One can clearly discard the class A probes, those that exhibit levels of hybridization that are not significantly above background. One can clearly distinguish also the class D probes. It is harder to demarcate the line between class B and class C probes. Both show higher-than-backgound and roughly equal hybridization signals in red and green channels. However, one can add an excess of unlabelled repetitive human DNA (also known as "Cot 1" DNA) to the hybridization mix to quench the hybridization signal to the probes that contain repetitive sequences, as is described in the literature (DeVries, S., et al., 1995, in Current Protocols in Human Genetics, ed Boyle, A. L.(John Wiley and Sons, Inc., New York), sup 6, unit 4.6, pp1-18). This unlabelled DNA serves as competitive inhibitor of the hybridization to the labeled samples. Thus any probe that shows diminished signal when excess unlabeled repetitive human DNA is used can be put into class B.

6.13. REPRESENTATIONAL PROBES AND REPRESENTED SAMPLES: DETECTING AMPLIFICATION

Although culling is useful to shape any collection of probes for the purposes of fabricating a more informative, and hence more efficient and economical, array, the protocol

described in the above example is not the best way to assemble a collection of useful probes. On average, for every 100 probes tested, only two become chosen as useful for a BglII represented sample.

A more efficient way to assemble a collection of probes useful for assaying represented samples is to select probes from similarly represented DNA. This DNA can be total genomic DNA from tissues or cultured cells, or genomic DNA that has been cloned as an insert into a cloning vector, such as a BAC or YAC, or cDNAs. In this way, the majority of the probes of the collection will share sequences with the represented sample, and each probe has a higher probability of being informative. Thus culling of these probes, after field testing, becomes a more efficient process.

In the following example, we make a BglII representation of total human DNA, from any source of normal cells, and individually clone the representation. (Different restriction endonucleases, or more complex representations, or even RDA products from megacloning vectors such as YACs (Schutte M., et al., 1995, Nucleic Acids Res 23(20):4127-33) or somatic cell hybrids could be used.) These probes are then arrayed and fall into two main classes: those that detect hybridization with repetitive sequences (class B) and those that detect hybridization with only single copy sequences (classC). The culling procedure is thus very efficient. Useful arrays can be fabricated with such probes even without removal of the class B probes, as they will comprise not more than half and probably about 30% of the total, and the addresses of such probes can be determined and recorded later.

We now fabricate an array with 10,000 probes, which is a very practical number. As in the Example 6.11, the arrays are hybridized with labelled BglII representations of tumor (green) and normal (red) DNA. As in Example 6.12, an excess of unlabeled repetitive DNA is added to quench hybridization from repeat sequences.

Now, instead of two probes detecting genetic gene amplification in the tumor (as indicated by statistically aberrant high green-to-red ratios), we will observe on the order of ten probes detecting amplification (0.1% of 10,000). Thus, even though the array with representational probes has one-tenth the addresses as the array used in Example 6.11, and is commensurably cheaper both to fabricate and analyse, it is five times more informative.

6.14. DETECTING GENETIC LOSS IN TUMORS

In the previous example we have illustrated the use of arrays hybridized to representations to detect gene amplification in tumors. A very similar protocol can be used to detect loss of genetic information in tumors. Such losses are often the hallmark of tumor progression, are usually indicative of genetic instability and the loss of a tumor suppressor gene, and can be used for diagnosis and prognosis of cancers. A variation in the protocol is necessary, because tumor biopsies invariably contain normal stroma, that is, normal cells such as fibroblasts, capillary endothelium, and blood cells. The DNA from these normal cells could obscure genetic loss within the tumor by common means of analysis, such as southern blotting and PCR analysis of loss-of-heterozygosity (Kerangueven F., et al., 1995, Genes Chromosomes Cancer 13(4):291-4; Habuchi T., et al., 1995, Oncogene 19;11(8):1671-4). It is therefore necessary to separate the tumor and normal nuclei.

Many tumors can be distinguished from normal cells, most commonly by aneuploidy (a different amount of DNA per nucleus) or surface markers. Hence in many cases tumor nuclei of tumor cells from a biopsy can be separated from normal stroma by fluorescence activated sorting (Del Bino G., et al., 1989, Anal Cell Pathol 1(4):215-23; Maesawa C., et al., 1992, Jpn J Cancer Res 83(12):1253-6) into populations that are 90% free of normal nuclei. Alternately, the normal stroma of tumor biopsy specimens can be microdissected and relatively pure populations of tumor cells obtained. DNA can

be prepared from as few as 5000 tumor cells or nuclei obtained by these means, and representations prepared. By comparing the tumor with normal representations in array format, as in the above example, genetic losses in the tumor can be detected.

This genetic loss can occur in two fundamental varieties. First, homozygous loss, where both copies of a gene have been lost in the tumor, will result in the absolute loss of those sequences. When those sequences encompass an element that both is present in the representation of the normal DNA and shares sequences with a probe of the array, the absence of those sequences will be detected by a high red-to-green ratio for that probe. That is, the array will detect the sequences present in the normal sample but absent in the tumor sample. We have estimated that, on average, a tumor loses about 3 megabases of sequence through homozygous loss, or about 0.1% of the genome. Thus we expect that about 10 probes from a 10,000 member array would detect loss.

Second, heterozygous loss, or LOH, can frequently be detected by an array based on representational analysis. In LOH only one of two alleles of a tumor is lost, and the explanation for the detection differs from that of homozygous loss. Individuals have genetic polymorphisms that are frequently manifest as restriction fragment length polymorphisms. Hence, sequences from one allele may be in a representation, due to being on a low molecular weight restriction endonuclease fragment, while sequences from the other allele are not. If the allele that is in a representation is the allele that is lost in the tumor, then that loss can be detected by the array, provided that sequence is shared with one of the arrayed probes. Previous estimates are that cancers lose about 15% of their genome through this mechanism. Depending on the density estimates of restriction endonuclease polymorphisms, upwards of 0.6% of the representation will be lost in the tumor, or about 60 probes per 10,000 member array.

6.15. OPTIMIZING HYBRIDIZATION CONDITIONS

Hybridization of nucleic acid, whether it be RNA or DNA, to the DNA fragments on the array is affected by several factors including complexity, concentration, ionic strength, time, temperature, and viscosity (Wetmur J.G., et al., 1968, Mol Biol 31(3):349-70; Wetmur J.G., 1976, Annu Rev Biophys Bioeng 5:337-61). By varying these factors we are able to optimize the hybridization conditions to allow for the highest signal with the lowest possible background.

By making a representation of the DNA sample we have already addressed the issue of complexity. In the event that the representation that we are using is still too complex to allow for favorable hybridization kinetics, we have the option of altering the representation to further decrease the complexity. One option to achieve this is to change the restriction enzyme that is being used for the production of the representation to an enzyme that cuts less frequently than the current enzyme we are using. A second option is to cleave the representation with a second restriction enzyme. However, the pitfall to doing this is that information is lost with every complexity reduction. Part of optimization entails choosing the representation that gives favorable hybridization kinetics but also yields as much information as possible.

The concentration of the sample also is an important factor that effects the rate of hybridization. The fact that we are producing representations for the hybridization puts us at an advantage as compared to many other chip techniques. We can make virtually unlimited amounts of representation for hybridization. In this way we can approach if necessary the maximum DNA concentration in solution. For example, sample concentrations of 1ug/ul up to 8 ug/ul can be used, if necessary.

Hybridization rates have been determined to be strongly dependent on the Na ion concentration ranging up to 3.2 M. One can, for example, start with a 0.5 M Na ion concentration, and vary this concentration from 0.25M up to

1M to optimize the Na ion concentration. Time of incubation also affects the completion of the hybridization reaction. One can vary this factor up until we reach 24 hours, or more if necessary. Preferably, we use the shortest time that will give us the best signal to noise ratio.

The temperature of hybridization may also be varied. The optimum temperature for hybridization of a fragment would be 25C below its melting temperature. We are asking for many fragments of different size and content from a representation to hybridize to their complementary probes in the microarray during the same incubation. Current protocols for hybridization use a temperature of 65C. One can vary this temperature from, for example, 55C to 75C to determine the optimum temperature of hybridization for our purposes.

The rate of hybridization can be increased by the addition of neutral polymers to the solution. It is believed that the polymer excludes water from solution increasing the local concentration of nucleic acid. One can increase the rate of hybridization by the addition of a neutral polymer such as ficoll.

6.16. MORE CULLING, LEADING TO MORE RELIABLE AND INTERPRETABLE ARRAY DATA

As will be understood by those of skill in the art, routine optimization of the arrays can take into account following. We have observed that a small minority of probes in an array are not reliable, in that they display variable hybridization signals even from representations prepared in parallel from the same sample. We presume that even in representations made as we have described, there may be some variability in the amplification of certain elements. It is useful therefore to test arrays made from a given collection of probes with multiple independent and parallel representations made from the same samples, to mark and note those probes which exhibit this behavior. They can then be culled from the collection.

We have previously discussed culling probes (Example 6.12), wherein useful probes are retained in a collection, and useless probes are discarded. This segregation of probes was physical, and arrays were fabricated that had higher concentrations of useful probes. However, when, as in this example, a small minority of probes is found to be useless, a more economical approach is achieved by making an electronic "black list", wherein the readings for a probe is marked as "to-be-ignored".

6.17. POLYMORPHIC ANALYSIS

As we indicated in Example 6.14, organisms are genetically polymorphic. That means that arrays based on representations can be used to provide a signature for individuals, which might be useful in forensic identification, or be used to follow genetic crosses between individuals, for example to determine paternity.

We can prepare representations, say BglII representations, from the DNAs of one individual, labeled with green, and compared to a "standard" human BglII representation labeled in red, using a BglII representational array containing, for example, 1000 probes. We estimate that polymorphic differences will be observed for roughly one out of 60 probes, resulting in differences (high green-to-red or high red-to-green ratios) at about 15 addresses out of 1000. This provides a "digit signature" for that individual as unique as a 1000 digit number in base three (with digits of green, yellow and red) with about 15 non-yellow digits. The number of such possible signatures is in excess of 10 to the 35th power. Larger arrays, or using compound representations of samples, as described in the next example, can provide an even more astronomically unique signature. From this genetic typing, that individual can be identified from a DNA sample.

The application of this method can be applied to a child, and the child's assumed biological parents, to determine if the parentage is correct. Due to the laws of Mendelian inheritance, if parentage is correct, all "green"

digits in the child's digit signature should have a green value in at least one parent. In more classical terms, the child possessing a "green allele" at an address (that is, the presence of the small fragment allele) must have inherited the same from either mother or father or both. Similarly, if the child displays a "yellow" digit for an address, then either mother or father must have a yellow digit at that address.

By comparison to different "standard" humans, this method of analysis can be further enhanced, because each comparison of different individuals will yield a different signature.

6.18. COMPOUND REPRESENTATIONS TO EXPAND LOH AND POLYMORPHIC ANALYSIS, AND ENABLE DETERMINATION OF POINT MUTATION LOADS

The simplest compound representation can be made by cleavage with a first restriction endonuclease, addition of linkers to those cleavage sites, cleavage with a second restriction endonuclease and then PCR amplification. This representation will consist of all the small fragments in the genome made by the first cleavage that do not contain restriction endonuclease sites for the second enzyme. By comparing samples made by compound representations, using representational arrays based on the first enzyme, we can thus score for polymorphic differences between samples at the second enzyme. Since the choice for the second enzyme is virtually unlimited, the same array can be used to detect many more polymorphic differences between two samples than can be detected without the use of compound representations.

Thus the use of compound representations expands the usefulness of arrays in the determination of identity (polymorphic analysis) and genetic loss in cancer (LOH analysis).

However, the use of compound representations makes possible a new use of representational arrays in cancer diagnosis. Cancers accumulate point mutations. Occasionally

these point mutations destroy a restriction endonuclease site. If the site destroyed is the site of the second enzyme, the compound representation of the tumor will contain a sequence that is not present in the same compound representation from the normal DNA of that patient. If the tumor representation is labeled in green and the normal in red, "green" addresses most likely will reflect point mutation in the tumor (after correcting for gene amplification, which can be determined by comparing the simple representations). This gives the tumor a digit signature of greens. The number of green digits reflect the point mutation load in the tumor, which may have predictive and prognostic value. Moreover, the signature of a biopsied tumor can provide a marker that can be used to determine if a second tumor arising in the same patient is an independent primary tumor or a metastasis of the first.

6.19. APPLICATION OF REPRESENTATIONAL APPROACHES TO EXPRESSION ARRAYS

The use of arrays of cDNA probes and cDNA oligonucleotide probes to measure expression levels is well established (Schena, et al., 1995, Science, 270, 467-70; Schena, et al., 1996, Proc Natl Acad Sci U S A, 93, 10614-9; Schena, M., 1996, Bioessays, 18, 427-31). In these uses, cDNAs or cRNAs from samples are prepared and analysed. The starting material from the sample is typically mRNA, and when the sample is available only in small amounts, it is problematical to perform the expression assays at all.

For this reason, it will often be desirable to prepare high complexity representations of the cDNAs prepared from limiting amounts of sample. Representations can be made in almost unlimited quantities from even small amounts of starting material, and therefore hybridized to chips at higher concentrations, thereby increasing the sensitivity and reliability of the expression assays.

In this example, it is not absolutely necessary that the probes of the array derive from the representation,

for the use of high complexity representations of the sample insures that most probes will share sequences with sequences amplified in the sample representations. Nevertheless, a cDNA probe array would function more efficiently if the probes are selected to share sequences with the representations made from expressed genes.

6.20. ARRAYS WITH COMPOUND DNA PROBES OR OLIGONUCLEOTIDE PROBES HYBRIDIZED TO REPRESENTATIONS

The above examples are described in relation to arrays of simple DNA probes. Each probe in the above examples comprises a single cloned sequence of DNA with a length roughly between 100 to 1000 bp (This range is not intended to define the term "simple DNA probes", it is merely an example thereof). The applications of arrays of compound probes, such as probes derived by representing YAC or BAC inserts, would not be much different. The major difference between arrays of simple probes and compound probes is that that LOH and polymorphic analysis could not be readily performed upon the latter. With arrays of compound probes gene amplification and homozygous loss could still be detected, essentially as described in Examples 6.13 and 6.14.

Another type of array can be made with oligonucleotide probes (Cho R.J., et al., 1998, Proc Natl Acad Sci U S A 31;95(7):3752-7; Pease A.C., et al., 1994, Proc Natl Acad Sci U S A 91(11):5022-6; Lipshutz R.J., et al., 1995, Biotechniques 19(3):442-7). There are advantages to fabricating such probes, both in terms of reproducibility, probe density, avoidance of repetitive sequences and cost (when large-scale production is desired). All the applications discussed in the above examples can be readily translated into the oligonucleotide format, provided that the oligonucleotide sequences of the array contained in the sequences of the representations used to prepare sample. Thus they will detect elements in the representation by hybridization.

This can be achieved in the following manner. DNA probes cloned from a representation are collected and sequenced. The sequencing does not need to be complete, and may extend merely as a single read from the ends of the cloning site. This sequence information is then used to synthesize the oligonucleotides that will be used on the array.

For some applications it may be preferable to first design arrays of simple DNA probes, and then characterize the properties of the probe arrays. Afterwards, the collection of DNA probes can be sequenced and the information used to format synthetic oligonucleotide arrays.

EQUIVALENTS

The foregoing specification is considered to be sufficient to enable one skilled in the art to broadly practice the invention. Indeed, various modifications of the above-described methods for biochemistry, organic chemistry, medicine or related fields are intended to be within the scope of the following claims. All patents, patent applications, and publications cited are incorporated herein by reference in their entirety for all purposes.

WHAT IS CLAIMED IS:

1. An array of DNA comprising probes derived from one or more representations of DNA.

2. The array of claim 1, wherein the DNA is genomic DNA.

3. The array of claim 1, wherein the DNA is from a megacloning vector.

4. The array of claim 1, wherein the DNA is from cDNAs.

5. The array of claim 3, wherein the megacloning vector specific sequences, and host specific sequences are substantially removed by representational difference analysis.

6. The array of claim 2, wherein the representation has a complexity equal to that of 70% or less of the complexity of genomic DNA.

7. The array of claim 2, wherein the representation has a complexity equal to that of 20% or less of the complexity of genomic DNA.

8. The array of claim 2, wherein the representation has a complexity equal to that of 5% or less of the complexity of genomic DNA.

9. A compound representation of DNA consisting essentially of those fragments of a first representation that do not possess a predetermined second restriction endonuclease site, said second restriction endonuclease site being different from that recognized by a first restriction endonuclease used to make the first representation.

10. A compound representation of DNA consisting essentially of portions of those fragments of a first representation that possess a second predetermined restriction endonuclease site, said second restriction endonuclease site being different from a that recognized by a first restriction endonuclease used to make the first representation.

11. A microarray being hybridized to one or more representations of DNA.

12. The microarray of claim 11, wherein two representations are being hybridized to the microarray.

13. The microarray of claim 12, wherein the two representations are compound representations.

14. The microarray of claim 12, wherein the representations are differentially labeled.

15. A method of producing a compound representation of DNA, comprising:

(a) digesting a sample of DNA with a first restriction endonuclease to provide digested DNA fragments;

(b) ligating an adaptor to said digested DNA fragments;

(c) amplifying said DNA fragments using primers complementary to said adaptors to provide a first representation of said DNA;

(d) digesting the first representation with a second restriction endonuclease to provide digested DNA fragments;

(e) amplifying said DNA fragments of step (d) using primers complementary to said adaptors to provide a compound representation of said DNA.

16. A method of producing a compound representation of DNA, comprising:

(a) digesting a sample of DNA with a first restriction endonuclease to provide digested DNA fragments;

(b) ligating a first adaptor to said digested DNA fragments;

(c) amplifying said DNA fragments using primers complementary to said adaptors to provide a first representation of said DNA;

(d) digesting the first representation with a second restriction endonuclease to provide digested DNA fragments;

(e) ligating a second adaptor to the ends of said fragments of step (d) created by the second restriction endonuclease;

(f) ligating a third adaptor to the 5' ends of said fragments of step (d);

(g) amplifying said DNA fragments of step (f) using primers complementary to the second and third adaptors to provide a compound representation of said DNA.

17. A method of hybridizing nucleic acids from one or more samples to an array of probe DNA immobilized on a surface of a solid phase comprising: contacting said array, containing or suspected of containing sequences complementary to nucleic acids from one or more samples, with nucleic acids from one or more samples, under conditions such that hybridization between the nucleic acids and probe DNA can occur, wherein the one or more sample nucleic acids are or are derived from representation(s).

18. A method of hybridizing nucleic acids from one or more samples to an array of probe DNA immobilized on a surface of a solid phase comprising: contacting said array, containing or suspected of containing sequences complementary to the nucleic acids from one or more samples, with the nucleic acids from one or more samples, under conditions such that hybridization between the nucleic acids and probe DNA can

occur, wherein the probe DNA is or is derived from one or more representations.

19. A method of hybridizing nucleic acids from one or more samples to an array of probe DNA immobilized on a surface of a solid phase comprising: contacting the array, containing or suspected of containing sequences complementary to the nucleic acids from one or more samples, with the nucleic acids from one or more samples, under conditions such that hybridization between the nucleic acids and probe DNA can occur, wherein the one or more sample nucleic acids are or are derived from representations and wherein the probe DNA is or is derived from one or more representations.

20. A method of producing a high complexity representation of DNA, comprising:

(a) digesting a sample of DNA with a relatively frequent cutting restriction endonuclease to provide digested DNA fragments;

(b) ligating an adaptor to said digested DNA fragments; and

(c) amplifying said DNA fragments using primers complementary to said adaptors to provide a high complexity representation of said DNA.

21. The method of claim 20 in which the sample of DNA is obtained from a biopsy specimen, a cell line, an autopsy specimen, a forensic specimen or a paleoentological specimen.

22. The method of claim 21 in which the biopsy specimen is a fractioned biopsy specimen or a microdissected biopsy specimen.

23. The method of claim 20 in which the sample of DNA is obtained from a fixed cell or tissue specimen.

24. The method of claim 20 in which the restriction endonuclease is selected from the group consisting of DpnII, Tsp509I, MboI, Sau3A1, MaeII, MspI, HpaII, BfaI, HinPI, Csp61, TaqI, MseI, AluI, BstUI, DpnI, HaeIII, RsaI, HnaI, and NlaIII.

25. The method of claim 20 in which the amplifying step (c) entails amplification by the polymerase chain reaction.

26. The method of claim 20 in which the DNA sample is a genomic DNA sample and the high complexity representation represents about 20-90% of the genome.

27. A method of producing a high complexity representation of DNA, comprising:

(a) digesting a sample of DNA with at least two restriction endonucleases to provide digested DNA fragments in which at least about 50% of said fragments are at between 100 and 1000 kb;

(b) ligating an adaptor to said digested DNA fragments; and

(c) amplifying said DNA fragments using primers complementary to said adaptors to provide a high complexity representation of said DNA.

28. The method of claim 27 in which the DNA sample is a genomic DNA sample and the high complexity representation represents about 20-90% of the genome.

29. Use of a high complexity representation for genetic analysis in a method selected from the group consisting of measuring sequence copy numbers by Southern blotting, measuring sequence copy number by quantitative PCR; monitoring loss of heterozygosity and monitoring genomic alterations through hybridization to chromosome spreads or panels of DNA probes.

30. Use of the high complexity representation obtainable according to the method of claim 20 or claim 27 for archiving a DNA sample.

31. Use of a the method of claim 17 to detect variations in gene copy number, variations in levels of gene expression, or the reproducibility of representations.

32. Use of a the method of claim 17 to detect variations in gene copy number, variations in levels of gene expression, or the reproducibility of representations.

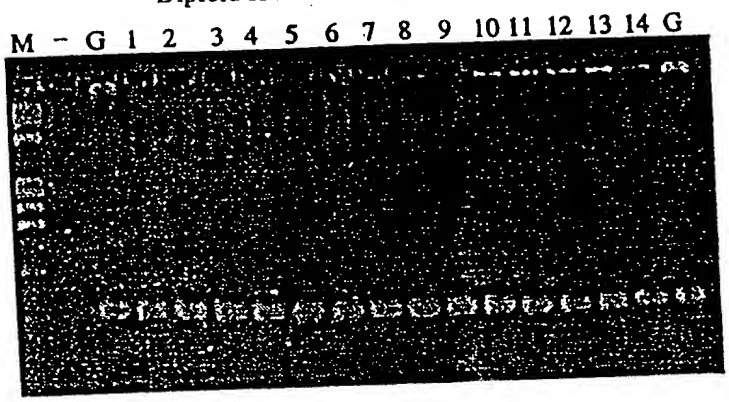
33. Use of a the method of claim 18 to detect variations in gene copy number, variations in levels of gene expression, or the reproducibility of representations.

34. Use of a the method of claim 19 to detect variations in gene copy number, variations in levels of gene expression, loss of heterozygosity, genetic polymorphisms, or to determine paternity or to map a sample nucleotide to a particular member of a megacloning vector library.

HCR Complexity

A

Diploid HCR from TumorBiopsies



B

WWW Probe

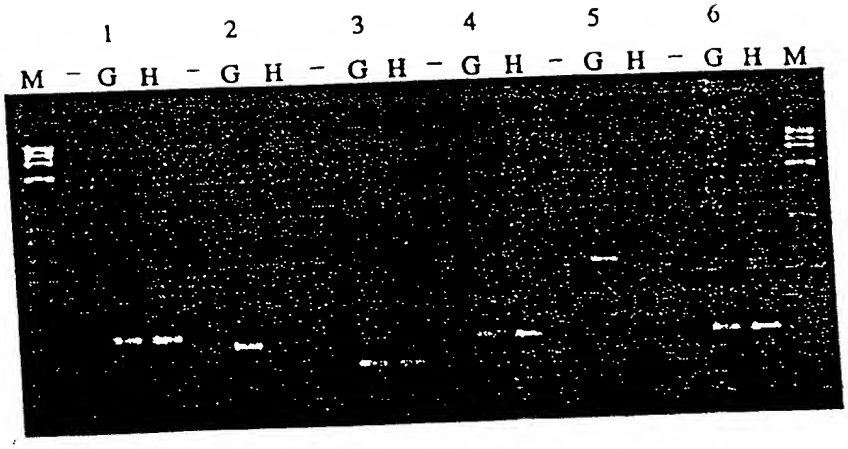
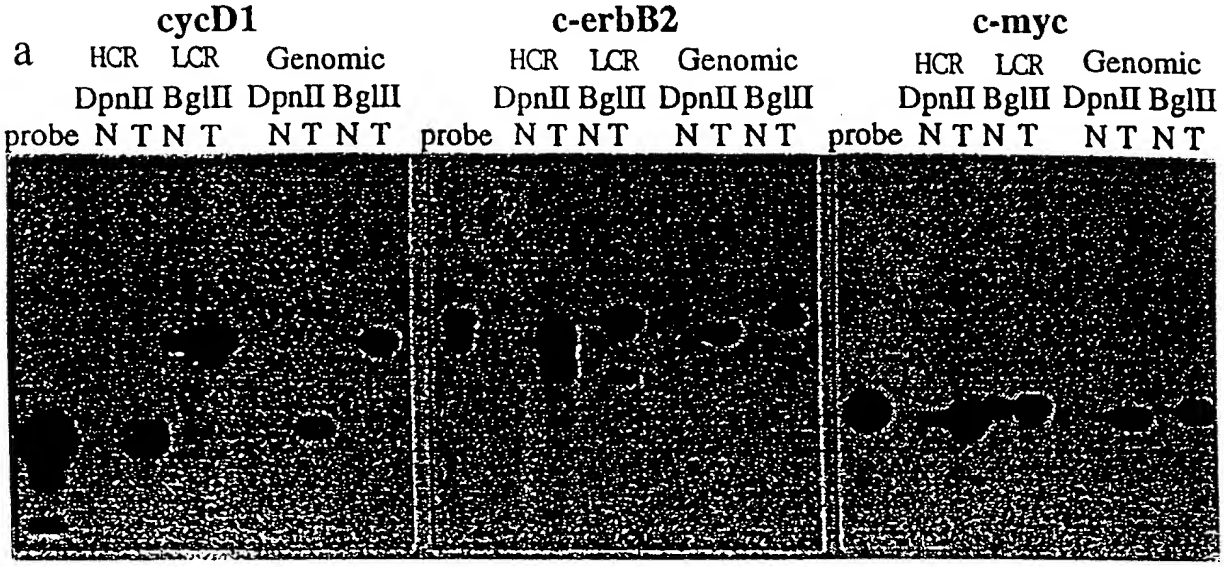


FIGURE 1

HCR: Quantitation of Copy Number

a



b

cycD1

	High DpnII	Low BglII
Representation	5.5	6.5
Genomic	5.1	6.4

c-erbB2

	High DpnII	Low BglII
Representation	9.2	11.3
Genomic	10.5	9.3

c-myc

	High DpnII	Low BglII
Representation	5.5	7.6
Genomic	5.2	5.2

Values represent the ratio of Tumor to Normal

FIGURE 2

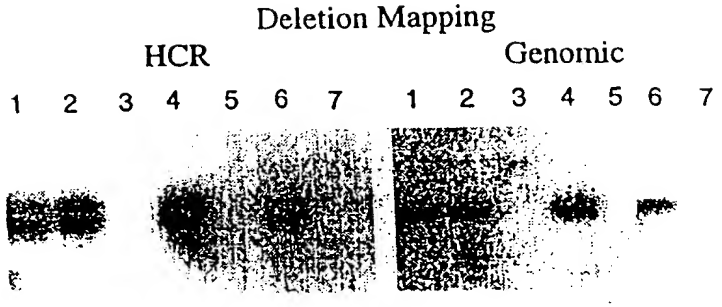
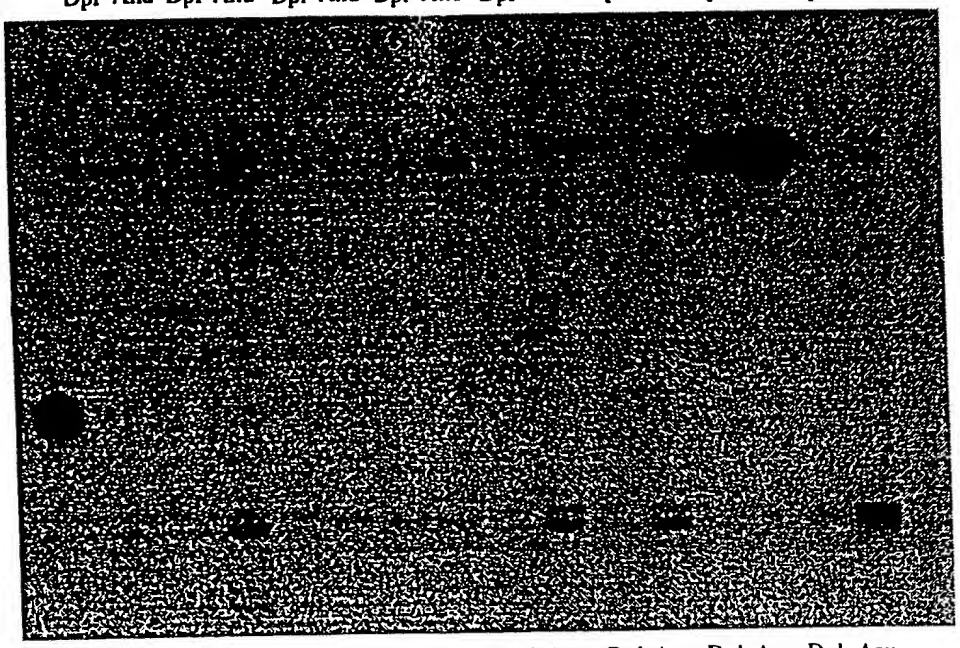


FIGURE 3

Primary Tumor HCR Panel

BBR3 BBR16 BBR33 BBR38 BBR44 BBR49 BBR50 BBR55

Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu



Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu Dpl Anu

BBR61 BBR66 BBR67 CHTN5 CHTN7 CHTN8 NSBR5 CHTN9

FIGURE 4

QPCR Analysis of HCR

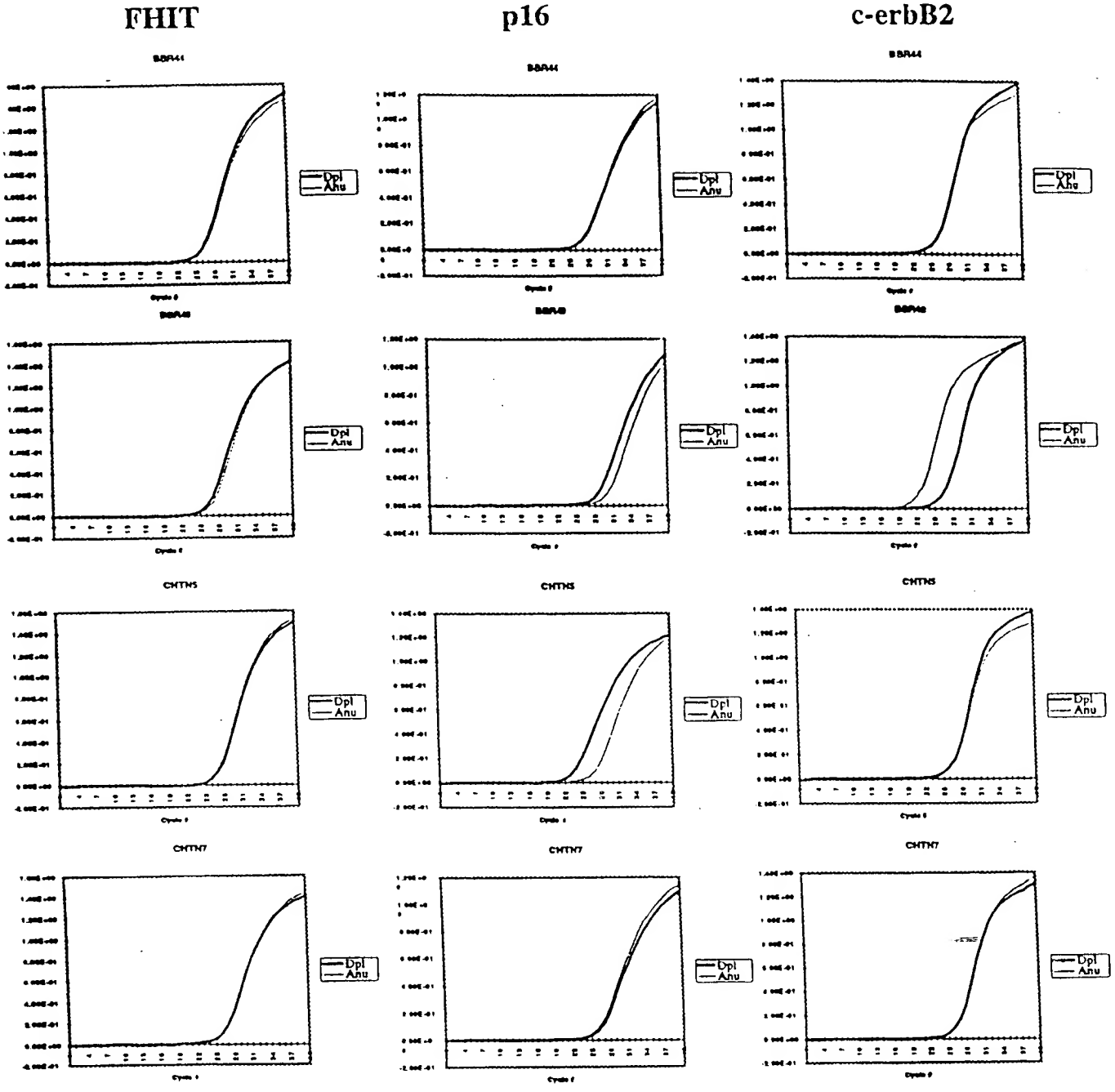


FIGURE 5

LOH Analysis of HCR

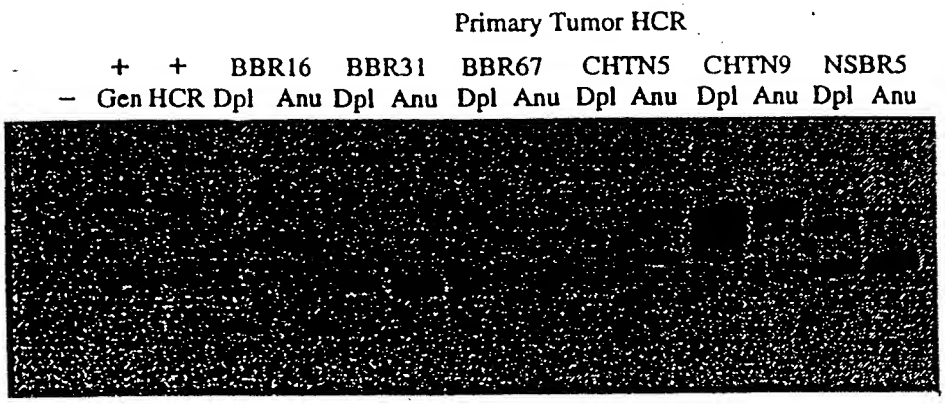


FIGURE 6

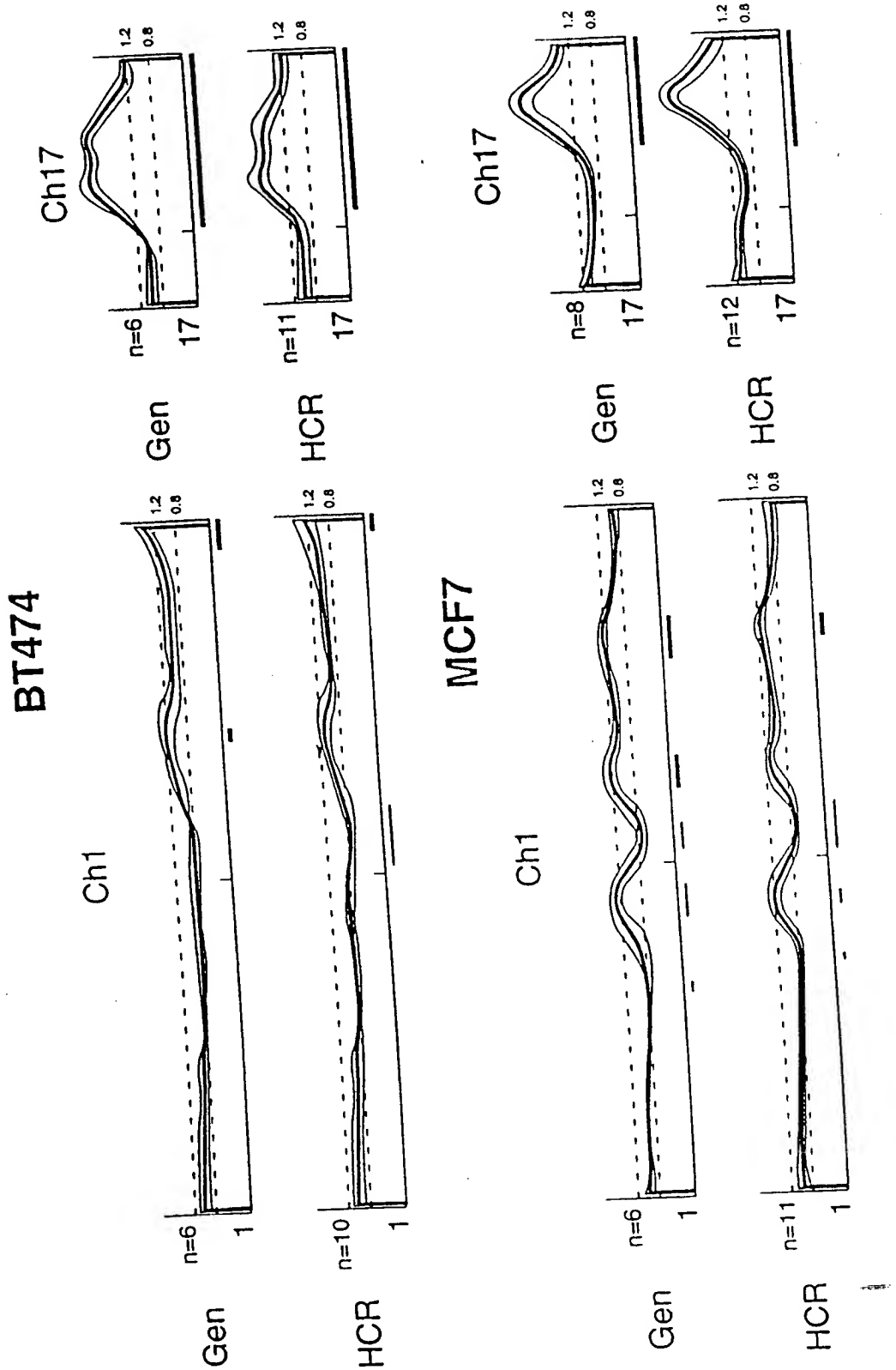


FIGURE 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/23168

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; C12P 19/34; C07H 21/04

US CL :435/6, 91.2; 536/24.3, 24.31

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 91.2; 536/24.3, 24.31

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT.

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,501,964 A (WIGLER et al) 26, March 1996 (03/26/96), see entire document,especially abstract and column4, lines 14-35.	1-34
Y	US 5,569,753 A (WIGLER et al) 29 October 1996 (10/29/96), see entire document, especially column 3, lines 35-50.	1-34
Y	US 5,510,270 A (FODOR et al) 23 April 1996 (04/23/96), see entire document, especially abstract.	1-8,11-14
Y	WO 89/10977 A1 (ISIS INNOVATION LIMITED) 16 November 1989 (16.11.89), see entire document.	1-8,11-14

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 15 JANUARY 1999	Date of mailing of the international search report 08FEB1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer Gary Jones Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/23168

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	GUO, Z. et al. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Research. 1994, Vol. 22, No. 24, pages 5456-5465, see entire document.	1-8,11-14
Y	LISITSYN, N. et al. Direct isolation of polymorphic markers linked to a trait by genetically directed representational difference analysis. Nature Genetics. January 1994, Vol. 6, pages 57-63, see entire document.	1-34
Y	HUBANK, M. et al. Identifying differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Research. 1994, Vol.22, No. 25, pages 5640-5648, see entire document.	3,4,32-34
T,P	GENG, M. et al. Isolation of Differentially Expressed Genes by Combining Representational Difference Analysis (RDA) and cDNA Library Arrays. BioTechniques. 05 September 1998, Vol. 25, No. 3, pages 434-438, see entire document.	1-34
T,P	CHANG, D.D. et al. Characterization of transformation related genes in oral cancer cells. Oncogene. 1998, Vol. 16, pages 1921-1930, see entire document.	1-34

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS,STN WPIDS,BIOSIS,MEDLINE,CANCERLIT,BIOTECHDS,LIFESCI,CAPLUS, EMBASE
rda,representational differentiation analysis,HCR, high complexity representation, probe, array, microarray, restriction, compound representation